Evaluating Explanation Performance for Clinical Decision Support Systems for Non-imaging Data: A Systematic Literature Review

Sneha Roychowdhury[1], Vita Lanfranchi[1], Suvodeep Mazumdar[2]

[1] School of Computer Science, University of Sheffield

[2]School of Information, Journalism and Communication, University of Sheffield

Abstract:

Purpose: This review investigates the effectiveness of Explainable AI (XAI) in machine learning (ML)-based clinical decision support systems (CDSS) using non-imaging data, focusing on explanation quality, clinical decision-making, user trust, and usability. It highlights clinician and patient perspectives to assess XAI's role in enhancing transparency and real-world adoption.

Method: A methodological and usability-focused systematic review using Web of Science, Scopus, IEEE Xplore, PubMed, Cochrane Library, and ACM Digital Library was conducted using keyword combinations, such as “XAI” AND “CDSS” AND “Evaluation metrics" AND "User study OR Evaluation study".

Results: The review identified an increase in multidisciplinary XAI healthcare research since 2023, with applications spanning intensive care, oncology, neurology, and clinical decision support systems. Studies commonly employed mixed-method evaluations, combining technical metrics (e.g., accuracy, fidelity) with human-centred assessments (e.g., trust, usability). Trustworthiness, interpretability, and transparency emerged as key XAI properties; however, aspects such as patient involvement, explanation usability, and clinical integration remain underexplored. Findings highlight ongoing challenges in balancing explanation faithfulness with user plausibility, and in aligning explanations with clinical reasoning and workflows.

Conclusions: The study highlights the importance of striking a balance between technical fidelity and human interpretability, achieved through human-centred evaluation frameworks that incorporate both objective and subjective metrics to enhance the real-world applicability of XAI tools. Future research should focus more on human-centred AI/XAI frameworks and real-world evaluations that prioritise multi-stakeholder collaboration to enhance clinical decision support, improve diagnostic accuracy, and enable personalised care without compromising clinician expertise or patient safety.

Keywords: Explainable AI (XAI), Healthcare, Clinical Decision support system (CDSS), Diagnostic accuracy, Personalised care, Objective evaluation metrics, Subjective or Human-Centred Evaluation metrics, Usability, Explainability, Interpretability, Trustworthiness.

1. Introduction

Machine learning (ML)-powered artificial intelligence (AI) methods are increasingly implemented in clinical decision support systems (CDSSs) to assist healthcare professionals (HCPs) in clinical decision-making, improve diagnostic accuracy and enhance personalised treatment for better patient outcomes and care experiences [1]. The integration of AI in healthcare has the potential to revolutionise patient care by enhancing diagnostic accuracies and streamlining clinical workflows, reducing delays in treatment and allowing clinicians to focus more on direct patient care. For example, De Fauw et al. (2018b) [2] highlights that AI-powered CDSS improve outcomes in retinal disease management by enabling accurate, timely referrals, supporting disease monitoring, and enhancing diagnostic precision. AI-powered CDSS are computer systems that help HCPs in decision-making by utilising AI techniques such as knowledge-based expert systems, ML, artificial neural networks (ANN), and genetic algorithms. It enhances CDSSs by processing and analysing large, complex datasets to improve diagnosis, treatment planning, resource allocation, and patient monitoring. However, AI-based CDSSs exhibit potential in laboratory settings [2,3] but due to challenges like "black box" opacity (e.g., ANN), non-diverse datasets and privacy and security concerns, governed by regulations like GDPR, HIPAA, and PIPL their real-world impact is often limited [4,5]. This can affect HCPs' trust and hinder HCP-patient-relationship, especially when AI recommendations deviate from clinical guidelines [4], raising medical, legal, ethical, and societal concerns, highlighting the importance of informed consent, liability, human-AI interaction, and trustworthiness when using AI-based CDSSs [6]. This emphasises how lack of explainability poses risks to core ethical values and public health [6] and the need to combine domain expertise with AI transparency [7] to ensure system trustworthiness [8] through informed validation, accountability by making system decisions traceable, and effective decision-making by aligning AI outputs with clinical reasoning. Explainable AI (XAI) refers to methods that make AI model decisions understandable to users by providing transparent, interpretable explanations especially in high-stakes domains like healthcare. In clinical decision support, XAI plays a critical role in helping HCPs validate recommendations, align outputs with clinical reasoning, and build confidence in the system. The evaluation of XAI methods must go beyond traditional performance metrics and include a multidimensional approach to be more effective in practice. This involves examining factors such as interpretability (how well users can understand the system's reasoning), usability (how easily the system's outputs can be applied in clinical workflows), and the balance between faithfulness (accurate explanations) and plausibility (meaningful explanations) ensuring that explanations are both technically accurate and meaningful to users [9]. To assess these aspects, both quantitative methods such as

fidelity scores and explanation similarity measures (e.g., cosine similarity, SSIM) [10] and qualitative methods including user interviews, think-aloud protocols, and observational studies [11] are required. Integrating these approaches enables a more comprehensive understanding of how XAI models perform in real-world settings and supports their safe, trustworthy, and ethically responsible integration into healthcare.

The growing reliance on AI in transforming the field of healthcare raises critical concerns about accuracy, ethical use, and the consequences of system errors or data breaches.XAI is essential in this context, enabling HCPs to validate decisions, communicate insights, and better understand model behavior [12]. However, explainability for non-imaging data remains significantly underexplored [13], despite its vital role in clinical decision-making and personalised care. Furthermore, there is a lack of systematic reviews addressing the usability, trustworthiness, and real-world applicability of XAI methods. This systematic literature review (SLR) is both novel and timely, offering a focused investigation into XAI for non-imaging healthcare data, highlighting multidimensional challenges like technical, practical, and human-centered, crucial for the successful deployment of XAI in healthcare. It employs a rigorous methodology to evaluate and synthesise existing research, addressing a clear gap in the literature.

The key contributors of this article are

a) It offers a SLR to gain original insights by synthesising evidence on the usability, trustworthiness, and real-world applicability of XAI tools, emphasising the crucial balance between faithfulness and plausibility to build appropriate trust.
b) It integrates both technical and human-centered perspectives, making it a valuable resource for researchers, clinicians, and policymakers trying to implement AI responsibly and effectively in healthcare.
c) It outlines potential future research directions to support the development and adoption of clinically relevant, human-centered XAI systems, thereby advancing the field towards more transparent and accountable AI applications in healthcare.

The paper is structured as follows: Section 2 provides an overview of the background of XAI in healthcare and reviews related works in this research domain. Section 3 describes the systematic review methodology covering research questions, search strategy, inclusion/exclusion criteria, study selection, data extraction, quality assessment, and threats to validity, ensuring transparency and reproducibility. Section 4 provides an overview and synthesis of previous studies on the effectiveness of XAI in healthcare, highlights study designs, user-centered approaches, and the challenges of real-world implementation and trustworthiness assessment through both qualitative and quantitative methods. Section 7 summarises key findings and concludes by outlining future

research directions to understand the potential of user-centered XAI in enhancing healthcare delivery and decision-making.

2. Background and Related Work

2.1. Explainable AI (XAI) in healthcare

XAI is an important subfield within AI that addresses the need for transparency and interpretability in AI models. It is important in healthcare to understand the rationale behind the AI decisions for widespread clinical adoption [14]. The use of XAI techniques across diverse data types in healthcare, including both imaging and non-imaging modalities has been increasing in recent times due to the extensive use of complex AI models. Imaging data such as X-rays, CT scans, and ultrasounds [15] has been the primary focus of XAI research, particularly in Computer-Aided Diagnosis (CAD) systems for conditions like cancer, where methods like saliency maps and Grad-CAM are commonly used to visualise regions contributing to a model's decision. In contrast, non-imaging data, including clinical measurements, medical signals , waveforms (e.g., ECG, EEG), and electronic health records (EHRs) , has received relatively less attention in terms of explainability, despite being extensively used in real-world CDSS [11]. For non-imaging data, deep neural networks, including models like CTGAN, medGAN, and TimeGAN, are effective for medical data synthesis but often suffer from overfitting and poor interpretability due to their data-driven nature. Techniques like Wasserstein loss can mitigate overfitting, and attention mechanisms may enhance explainability, though they are underexplored in non-imaging contexts. In contrast, knowledge-driven methods such as EMERGE and Bayesian Networks support expert input but struggle with high-dimensional data and can be time-consuming to implement. Hybrid approaches, such as theory-driven modeling and integrating prior knowledge into deep networks, are emerging as promising solutions to improve both synthesis quality and interpretability. This highlights the importance of integrating expert knowledge,a component of user-centered design [11]. Within non-imaging data, structured data (e.g., lab values, vital signs, demographics) is often considered inherently interpretable, and some researchers argue that interpretable models such as decision trees or logistic regression may be preferable, especially when the marginal performance gains of complex black-box models do not justify the loss of interpretability [16], hence the need of post-hoc XAI methods are not needed. Despite this , there has been increasing use of ML/DL methods with XAI for tabular EHR data [17] (2017–2023), where SHAP emerged as the most widely used method for its theoretical robustness and broad applicability. Existing literature shows that explainability methods for tabular data are predominantly model-agnostic, often designed to interpret complex models like Random Forests or

Gradient Boosting. Common methods include feature ablation, permutation importance, and impurity-based scores. Advanced tools like SHAP and LIME offer both local and global insights, with extensions like ALIME and Anchors improving reliability. Counterfactual explanations show how small input changes could alter a model's decision. Sensitivity analysis and visual tools like PDP, ICE, and ALE help understand feature effects. Rule-based models like InTrees can approximate black-box behavior . For time series, XAI methods are often adapted from computer vision. Gradient-based techniques (e.g., saliency maps, Grad-CAM) and perturbation methods (e.g., Occlusion) . Attention mechanisms in recurrent neural networks (RNNs) and transformers naturally enhance interpretability. Other tools like SAX, fuzzy rules, and Shapelets extract key patterns to explain model behavior over time [15]. In the literature it has been also underscored that RNNs for sequential or temporal data and convolutional neural networks (CNNs) with temporal convolutional layers have become state-of-the-art for time series tasks like classification, forecasting, and clustering since they improve accuracy and can avoid heavy preprocessing required by traditional methods [18]. This underscores a critical need to synthesise literature on evaluation of XAI methods in terms of their effectiveness, usability, trustworthiness and real-world applicability for non-imaging data so as to enhance the research towards more clinically meaningful, trustworthy, and actionable decision-making .

2.1.1. Core Concepts: Interpretability, Explainability, Transparency, Accountability, and Trust, Trustworthiness, Faithfulness , plausibility and Usability

Interpretability, transparency, accountability, and trustworthiness are widely recognised as key principles of XAI in healthcare [19]. While the terms interpretability and explainability are often used interchangeably, they have conceptual differences: interpretability refers to the extent to which a human can understand and predict an AI model's behavior without requiring deep technical expertise, whereas explainability focuses on how clearly the model's decision-making process is communicated [20]. These concepts underpin transparency and support accountability, which is essential for identifying and correcting errors, biases, and unintended consequences [19]. Trust and trustworthiness, although closely linked, are not synonymous; trust is a psychological state of users' confidence in a system while trustworthiness refers to the actual, evidence-based reliability of that system [12]. In practice, XAI seeks to balance interpretability and predictive performance, either through inherently interpretable models (e.g., decision trees) or through post-hoc explanations of black-box models [20]. However, a critical challenge arises when seemingly intuitive explanations fail to accurately reflect the model's internal logic. Studies have shown that models may produce plausible but misleading explanations, which can obscure spurious correlations or irrelevant features, potentially leading to over-reliance on the AI system . This

highlights the need for faithfulness, the degree to which an explanation accurately represents the model's reasoning. Unlike plausibility, which is often evaluated through human judgment, faithfulness should be assessed independently of user perception to ensure objective reliability . Even interpretable models must be rigorously tested for faithfulness since an intuitive output can still mislead [13]. Finally, the usability of explanations is defined as how they are presented and how effectively they support user understanding which is crucial for fostering trust and confidence. Interactive, user-centered explanation designs have been shown to reduce uncertainty and increase confidence in AI systems, reinforcing the need for human-centered, trustworthy XAI in clinical contexts [20]. Usability, faithfulness, plausibility, and trustworthiness in XAI often create trade-offs. For example, plausible explanations may misrepresent the model, undermining faithfulness and trustworthiness, while faithful explanations can be too complex to understand, reducing usability and user trust. Thus designing explainable systems not only require optimising individual dimensions , but also balance the trade-offs that emerge between them . This emphasises effective explanation strategies that must adapt to context, user role, and domain risk while carefully balancing the trade-offs between faithfulness, trustworthiness, plausibility, and usability.

2.1.2. Evaluation of XAI

The evaluation of AI models involves performance and explainability as key assessment criteria. Traditional AI evaluation focuses on how well the model performs in terms of making accurate predictions, generating high-quality outputs, or classifying images etc. using metrics like accuracy, precision, recall, F1 score, error rates, ROC\AUC and Kappa statistics [23]. In contrast, explainability ensures that AI decisions are interpretable and trustworthy, aiming to make the system's behaviour transparent. The quality of explanations in XAI is assessed through both objective, computer-centred evaluation methods and subjective, human-centred evaluation methods [24]. Objective methods include metrics such as fidelity, completeness, monotonicity, and other performance metrics like D (Performance Difference), R (Number of Rules), F (Number of Features), S (Stability) etc., which assess the performance and complexity of the explanation model [24,25]. Furthermore, these metrics are divided into model-based explanations (model size, model complexity etc.), attribution-based explanations (monotonicity or sensitivity etc.) and example-based explanations (non-representativeness, diversity etc.) [25,26 27']. In contrast, human-centred methods focus on evaluating XAI explanations based on their usability, understandability, and trustworthiness from the user's perspective . These methods gather feedback to assess how easily users can interact with the system, comprehend the explanations, and apply them effectively to ensure AI's reasoning is clear, trustworthy and helpful as well as user-friendly, promoting ethical and practical

application in real-world clinical settings [27 26']. The subjective XAI metrics include transparency, trustworthiness, effectiveness, satisfaction, explanation goodness, user curiosity/attention engagement, user understanding, user performance/productivity, system controllability/interaction, explanation usefulness, interactivity, interestingness, informativeness, human-AI task performance etc. [25]. Effectively assessing explainability to improve model transparency poses a significant challenge due to the lack of a unified standard or consensus among researchers [28]. This lack of agreement on fundamental properties, criteria or approaches for explainability in AI complicates efforts to establish a universally standardised framework . Hence to address this, Ali et al. (2023) [20] proposed a four-axis framework covering data, model, post-hoc explainability, and explanation assessment. Data explainability ensures data quality through tools like Google Facets and techniques like EDA and knowledge graphs. Model explainability addresses the transparency of AI models through interpretable models, hybrid approaches, and methods like TED, RNP, and regularisation techniques. Post-hoc explainability involves interpreting model predictions after training using methods like Deep Taylor Decomposition for feature attribution, visualization techniques such as Grad-CAM, example-based approaches like LIME, and game-theoretic tools like Shapley values. It also includes knowledge extraction methods like model distillation, which simplify complex models to improve understanding. Finally, assessment of explanations measures explanation quality based on fidelity, comprehensibility, fairness, and user satisfaction, ensuring alignment with human decision-making and ethical standards [20]. This is comprehensive but its broad scope raises concerns about practical applicability, particularly in clinical settings where time, interpretability, and integration constraints are critical. Techniques like SHAP, LIME, and model distillation offer valuable insights, yet they often trade off between faithfulness and usability. Moreover, the assessment dimensions fidelity, fairness, comprehensibility, and user satisfaction lack consistent operationalisation, making comparative evaluation difficult. As such, despite its ambition, the framework underscores the persistent challenge of translating theoretical explainability into actionable, context-specific XAI systems. To address these challenges, frameworks should focus on context-specific, user-centered explanations, standardise evaluation metrics, balance simplicity with fidelity, and involve end-users in design and testing.

2.1.3. User-Centered Explainability

Multidisciplinary collaboration among developers, HCPs, and policymakers is essential for effective integration of AI into clinical workflows to ensure the development of transparent, ethical, and user-centered XAI systems [6]. A key aspect of this integration is increasing transparency through explainability, which fosters trust and confidence.

However, its effectiveness depends on factors such as technical feasibility, clinical context, its role in decision-making, and the end-users [4]. For instance, physicians who understand AI decision-making are more likely to trust its recommendations and question them when necessary, improving decision-making and strengthening trustworthiness in the AI-user dyad. This highlights where human expertise and AI interaction are critical, explainability is key to fostering trustworthiness and effective outcomes. Hence, it is essential to tailor it to specific contexts and user needs to maximise AI's effectiveness in healthcare settings. Different stakeholders, clinicians, patients, healthcare administrators, legal entities and developers, have distinct requirements for AI explanations. For instance, HCP's requires accurate and trustworthy XAI systems which have accessible user interfaces and can be seamlessly integrated into existing clinical workflows providing actionable insights , patients require clear risk assessments, treatment options, and reasoning behind AI-generated recommendations emphasising the need for simple, understandable, personalised and transparent explanations and developers need transparency for model debugging, bias correction, and performance optimisation [2,7]. This underscores the importance of tailored explanations. The study [21] highlights that user-centered explainability involves developing explanations that are specifically aligned with the roles, needs, and expertise of clinical end-users, such as physicians and nurses. Barda et al. (2020) [21] developed a framework for designing explanations informed by end-user needs for the pediatric intensive care unit (PICU) mortality risk model. The study used shapley values for instance-level, model-agnostic explanations and refined explanation displays through focus groups with critical care nurses and physicians. Feedback highlighted the importance of minimising cognitive effort and tailoring explanations to diverse clinical roles and expertise levels. The final result of user-centered display incorporated these insights positively, supporting the use of transparent, interpretable explanations to enhance physicians understanding and acceptance of ML predictions. This work advances effective communication of ML model information in clinical settings and underscores the importance of involving users in the design of AI explanations [21]. User-centered design is essential for knowledge-informed ML systems in healthcare. This enables effective integration of diverse domain knowledge from disease information to expert clinical insights into explanations that genuinely align with end-users' needs and understanding.These explanations align better with medical reasoning, but many remain technical and developer-focused, limiting their accessibility for clinicians with varied expertise and patients. Hence, to maximise their impact, explanations are tailored to fit users' backgrounds, workflows, and cognitive needs. This ensures they are practical and usable within time-constrained clinical environments. Moreover, patients, who are directly affected by medical decisions, are often neglected in explainability efforts despite needing clear and transparent information. This is important because aligning

explanations with how HCPs communicate and considering all stakeholders needs builds trust and understanding [22].

2.2. Review of prior systematic studies of XAI in healthcare

In recent years several systematic reviews have examined the role of XAI in healthcare. This reflects a growing interest in ensuring transparency, interpretability, and trust in clinical AI applications. These reviews have typically focused on examining the technical and methodological aspects of XAI in healthcare, such as identifying and categorising commonly used algorithms, analysing datasets and performance metrics, and outlining domain-specific challenges like bias, complexity, and regulatory constraints. While they offer valuable insights into model-level explainability and system-level implementation, they tend to underemphasize user-centered aspects particularly how explanations are perceived, interpreted, and utilised by end-users such as HCPs and patients. Consequently, few studies rigorously evaluate key user-focused factors such as usability, cognitive fit, and trustworthiness of XAI outputs in real-world clinical contexts, revealing a critical gap in translating technical advances into effective, user-friendly tools. Additionally, limited attention has been given to the needs of non-expert users or to tailoring explanations based on user roles and settings. As a result, there is a clear gap in understanding the human factors that influence the acceptance, effectiveness, and impact of explainability in healthcare practice. This review aims to address key gaps by systematically evaluating the effectiveness of XAI methods across various healthcare domains. It focuses on user-centered approaches, clinical usability, trust-building dimensions, and real-world applicability maintaining a nuanced balance between faithfulness and plausibility. To comparatively analyse the prior systematic reviews methodological critique and thematic analysis is conducted as follows.

2.2.1 Methodological Critique

The SLRs for XAI in healthcare exhibit generally rigorous and structured methodologies but there are several limitations affecting transparency, comprehensiveness, and reproducibility. Many reviews adopt established frameworks like PRISMA, PRISMA-ScR, or Kitchenham & Charters, but often lack consistency. Ambiguous review classification, lack of formal quality appraisal, insufficient detail in data extraction and synthesis, and missing inter-rater reliability or protocol registration (e.g. [29,31,32, 34, 36, 37, 42]) are some of the common issues identified. Some reviews restrict scope by excluding non-Q1 journals or conference papers, limiting inclusivity [30]. Moreover, some struggle with methodological clarity, opaque filtering, and

inconsistent application of criteria like PICO [32]. Although a few SLRs demonstrate procedural rigor and broad database coverage [34, 40], many rely heavily on single frameworks or lack robust multi-reviewer participation, weakening reliability [33, 35].Reviews that attempt broader integration of technical and socio-ethical perspectives often lack clarity in inclusion criteria and data transparency [37]. Overall, while some SLRs show methodological robustness [40], key gaps remain in standardised quality appraisal, detailed synthesis methods, and reproducibility measures, reducing the overall credibility and replicability of findings in this interdisciplinary domain. Therefore, while all reviews contribute valuable insights to the evolving field of XAI in healthcare, they vary considerably in scope, methodological appropriateness, and transparency. Only a few, most notably Eke and Shuib (2024) [38] and Islam et al. (2022) [40] demonstrate robust procedural planning and execution aligned with recognised guidelines.

2.2.2 Thematic Analysis

a) Healthcare Application Areas

There has been a significant increase in XAI research, with 54% originating from the healthcare sector, emphasising explainability, interpretability, and trustworthiness as dominant properties and highlighting the growing focus on transparency and accountability in AI-driven clinical applications in a comprehensive analysis from 2018 to 2023 [39] . XAI is most prominently applied in disease diagnosis, which includes cancer (lung, breast, colorectal, skin), Alzheimer's, brain tumors, and cardiovascular conditions. Diagnosis has been a dominant use case due to its high clinical risk and the abundance of imaging data [29-34,38,41,42]. In prognosis and risk prediction, XAI supports models estimating survival rates (e.g., breast cancer), mortality risk (e.g., ICU or COVID-19), and disease progression (e.g., brain aging, stroke recovery), improving transparency and feature-level insight [29-31,34,35,38,40,42]. For treatment decision support, explainable systems assist in personalised therapy planning, such as surgical options or drug therapy recommendations for breast or laryngeal cancer allowing clinicians to validate model reasoning against clinical guidelines [29-35,38,40]. Emerging applications include clinical pathway prediction for forecasting sequences of medical events using EMR data and patient-centered tools like fuzzy logic-based self-monitoring apps , which empower patient-centered care [30,32,34,35,41]

b) XAI Methods and Techniques

Model-agnostic post-hoc methods are the most frequently used, particularly SHAP , LIME, Anchors, and rule extraction methods. These tools explain model predictions after training and are applicable to diverse data types [29–32,34,35,39]. Interpretable-by-design models, such as

decision trees, fuzzy logic, and Bayesian networks, provide built-in transparency. These are favored when model logic must be directly understandable, such as in regulatory or ethical contexts [32,35,38,40,41] . Deep learning explanation techniques, including CAM, Grad-CAM, saliency maps, DeepLIFT, and integrated gradients, are essential for image-based tasks. Attention mechanisms and spatial attention maps are also used for text, image, and time-series data [30,31,33-35,37,39]. Hybrid methods combine symbolic reasoning (e.g., Bayesian models) with neural networks for richer, multilevel explanations. An example is MulNet, which jointly processes EMR and imaging data to deliver interpretable outcomes [31,35,37,39]

c) Types of Healthcare Data

Medical imaging (MRI, CT, X-ray, ultrasound, dermoscopy) dominates in XAI research, especially when paired with deep learning and visual explanations like heatmaps or activation maps [29–32, 34, 38, 41, 35]. SHAP and interpretable models are widely used for EMR, structured clinical data. These datasets often include demographics, lab tests, and vital signs [29, 30, 31, 34, 37, 39, 35]. Biosignals such as ECG, EEG, and heart rate are widely used in monitoring applications, requiring specialised temporal models and tailored XAI techniques to handle their time-series nature [30, 31, 34, 35]. Omics data including genomics and proteomics supports precision medicine and is typically interpreted using feature attribution methods like SHAP and integrated gradients [31, 34, 37, 39, 35]. While less common, clinical notes and unstructured text data are an emerging focus, with attention-based NLP models providing meaningful explanations for complex textual information [29, 30, 34, 39, 35].

d) **Clinical Use Cases**

Disease detection such as cancer, cardiovascular diseases, Alzheimer's, and brain disorders is the key use case for XAI where transparency is critical for clinician trust. These tasks often rely on imaging data and post-hoc visualisations [29–31, 33, 34, 38, 41, 35]. XAI plays a role in COVID-19 diagnosis and prognosis, helping to predict ICU admissions, disease severity, or mortality using imaging, symptoms, or lab data [30, 34, 38, 35]. XAI supports risk stratification models by validating the rationale behind patient grouping for outcomes like surgical complications or cancer recurrence, which is essential for fairness and trust [31, 33, 34, 38, 35]. In therapeutic guidance, it helps explain treatment recommendations such as laser surgery eligibility, laryngeal cancer therapy plans, or drug dosage adjustments [31, 34, 38, 40, 35]. Additionally, XAI is applied in medical education and decision support interfaces, where visual or interactive explanations enhance understanding for both junior clinicians and patients [32, 33, 38, 41, 35].

e) **Evaluation Metrics for XAI**

Human-centered evaluation and Functionality-grounded evaluation have been identified as the Evaluation Metrics for XAI for these SLRs. Human-centered evaluation focuses on user

perceptions such as trust, confidence, satisfaction, and perceived usefulness usually measured via surveys, interviews, and expert panels [31, 33, 40, 41, 35]. Functionality-grounded evaluation uses quantitative metrics like fidelity (consistency with model behavior), accuracy, completeness, and compactness of explanations. Some studies also evaluate agreement between model explanations and expert decisions [33, 39, 40, 35]. Explainability properties assessed in XAI include local versus global explanations, various interfaces such as visual, textual, or interactive, and reasoning types like contrastive or contextual. While measures of comprehensibility and clinical relevance are crucial, they are often subjective [31, 33, 39, 40, 35]. Despite these efforts, most reviews emphasise the lack of standardised evaluation frameworks, with the field relying largely on qualitative feedback or indirect metrics like diagnostic accuracy, which limits rigorous and consistent comparison of XAI methods [33, 39, 40, 41, 35].

f) **Reported Limitations and Gaps**

There are several limitations that have been identified from these SLRs. A major limitation is the lack of human-in-the-loop design since most XAI systems do not involve physicians during development or evaluation, reducing clinical relevance and trust [31, 33, 34, 40, 41, 35]. Limited real-world deployment is also an emerging issue because many studies remain retrospective or lab-based, lacking operational validation in hospital environments [32, 34, 37, 41, 35]. Other limitations include poor evaluation of explanation quality, user interface design, accuracy-interpretability trade-off, regulatory and ethical gaps. Poor evaluation of explanation quality remains, as there are few standard metrics and minimal consensus on what constitutes a "good" explanation in clinical settings [33, 39, 40, 41, 35]. User interface design challenges persist because few systems support interactive, contextual, or socially relevant explanations suited to clinical workflows [32, 33, 38, 41, 35]. The accuracy-interpretability trade-off is especially prominent in deep learning systems, where complex models often outperform interpretable ones but at the cost of transparency [31, 32, 39, 35]. Lastly, regulatory and ethical gaps remain. Many systems lack compliance with GDPR or FDA requirements, and explainability alone is not always sufficient for legal or ethical accountability [32, 37, 41, 35].

Table 1: Comparison of prior systematic review studies.

Study	Year of Publication	Research aim	Years of publication included	Databases searched	Number of studies preliminarily identified	Number of Final studies included

					before exclusion	
[29]	2023	Identify the most common XAI algorithms, methods, and tools utilized by researchers in these fields. Examine the challenges and limitations encountered in XAI research for medical and healthcare applications. Determine the datasets that are most prominently used in XAI research within these domains. Analyse the key performance metrics considered in evaluating XAI models in medical and healthcare research	2018-2022	IEEE Xplore, ACM Digital Library,Scopus , Web of Science	454	93
[30]	2022	To provide areas of healthcare that require more attention from the XAI research community.	2018-2022 (March)	IEEE Xplore,PubMed,Scopus , Web of Science	1194	99
[31]	2024	XAI methods identification and categorisation XAI literature review with special focus on healthcare domain Ascertainment of XAI challenges and problems in healthcare.	2017-2024	Elsevier Springer Taylor & Francis Semantic Scholar ACM Digital Library IEEE Xplore	324	113
[32]	2023	Explore the role of XAI in healthcare and medicine. Identify the limitations of standard AI in addressing critical challenges in medical	Jan 1, 2012 – Feb 2, 2022	Elsevier (ScienceDirect), IEEE, MDPI, Springer Nature,	17300	110

		applications. Determine the most effective methods for achieving interpretability in AI-driven healthcare systems. Establish the optimal timing for implementing XAI in different stages of medical AI development and deployment.		Bentham Science, Wiley, Medrxiv, ArXiv, IEEE, ACM etc		
[33]	2024	To assess the role of HCI evaluation techniques in achieving XAI goals. To examine the evaluation methods employed for XAI goals in HCI. To explore the implementation of XAI goals. To ascertain the emphasis placed on specific goals in research. To examine the domains where XAI goals are applied. To review the ML models utilized for various XAI goals. To identify the challenges in implementing XAI goals across different fields.	2018-2023 (Jan 21)	ACM digital library, Scopus, SpringerLink, IEEE	4295	101

[34]	2024	To investigate the adoption of Explainable AI(XAI) methods across specific healthcare domains. To identify and analyse the types of datasets commonly used in explainable CDSSs and their key characteristics. To examine the prevailing trends and most effective ML models used in explainable healthcare systems. To explore the most commonly utilized XAI methods in electronic healthcare and their contribution to the interpretability of ML models. To understand how multiple XAI methods collectively contribute to ensuring explainability in the healthcare decision-making process.	2000-2023	Cochrane Library, PubMed, Scopus, Web of Science	1226	63
[35]	2023	Identify XAI techniques used in medical applications. Explore the relationship between XAI techniques and medical tasks. Evaluate the usefulness of explanations from a physician's perspective. Assess the potential for clinical application of XAI solutions. Address explainability challenges and identify	2019-2022 (Oct)	IEEE Xplore, PubMed, Scopus , Web of Science, Other sources	769	198

		areas for further development.				
[36]	2024	Explore challenges in implementing XAI in healthcare. Evaluate the effectiveness of XAI in improving clinical decision-making. Analyse the XAI algorithms used in healthcare applications.	2020-2024 (Mar 5)	Google Scholar and Pubmed	50	40
[37]	2024	Role of taxonomies and classifications in AI assessment and integration in healthcare. Support for AI in safety-critical healthcare systems. Identification of tools for transparency, interpretability, and decision-making. Evaluation of their effectiveness in healthcare applications. Influence on HCPs decision-making and patient trustworthiness. Classification and implications of XAI techniques in healthcare. Review of datasets used in interpretable AI models and their real-world representativeness. Identification of biases and challenges in transparency-focused	Jan 1,2014-Jan 1,2024	PubMed, Springer, IEEE Xplore, Scopus, ACM Library, ScienceDirect, Google Scholar, and Web of Science.	1837	148

		datasets. Current open research challenges in AI transparency and interpretability. Potential future research directions to enhance AI trustworthiness and ethical considerations in healthcare.				
[38]	2024	Identify prevalent approaches and techniques used to enhance explainability and transparency in AI healthcare systems. Examine commonly employed ML/DL techniques in XAI for healthcare. Analyse datasets frequently used to train ML/DL models in XAI for healthcare. Identify key performance metrics used to evaluate ML/DL models in XAI for healthcare. Investigate validation methods commonly considered during the training of ML/DL models in XAI for healthcare.	2015-2023	Scopus, IEEE Xplore, Science Direct, Wiley, and Springer Link	567	69

[39]	2024	To identify and analyse sectors where XAI improves trustworthiness, transparency, and decision-making. To explore key properties of XAI across domains, identifying essential factors like fairness, transparency, and accountability. To evaluate the effectiveness of existing XAI algorithms and frameworks, assessing their self-explanatory nature and impact. To identify and overcome domain-specific challenges in XAI implementation, focusing on regulatory, data, and trustworthiness issues.	2018-2024	IEEE Xplore Digital Library, ScienceDirect, SpringerLink, ACM Digital Library, Google Scholar, ProQuest, and PubMed.	298	53
[40]	2022	Identify and categorise application domains and tasks where XAI is applied. Analyse and classify XAI methods used in different domains and tasks. Explore different forms of providing explanations (visual, textual, interactive, etc.). Identify and analyse evaluation metrics for XAI methods across domains Identify and categorise the research topics in XAI literature, with focus on healthcare. Examine methodologies and protocols used in SLRs on XAI.	2028-2021(June)	Google Scholar, SpringerLink, IEEE Xplore, ACM Digital Library	1709	137

[41]	2024	Explore the limitations in current XAI research and identify areas for future work. Provide an overview of current explainability techniques in XAI. Investigate the benefits of using explanations in XAI, such as transparency and trustworthiness. Analyse the negative impacts of explanations in XAI, including bias and system complexity. Examine the evaluation methods for XAI explanations and identify gaps.	2019-2023(May)	ACM Digital Library, IEEE Xplore, and ScienceDirect.	51	29
[42]	2023	Investigate applications of deep learning, federated learning, and XAI in healthcare and disease management. Analyse forums, venues, and trends in publications on deep learning, federated learning, and XAI in healthcare.	2015-2024	The ACM Digital Library IEEE Xplore Digital Library Google Scholar ScienceDirect SpringerLink	419	24

2.3. Research gaps

This SLR is motivated by several critical gaps observed in prior reviews of XAI in healthcare:

a) Limited focus on human-centric aspects: Existing SLRs primarily emphasise on technical performance, with insufficient attention to human-centric factors such as stakeholder trust, usability, and user experience in clinical workflows.

b) Lack of multi-stakeholder perspectives: There is a scarcity of studies that consider diverse perspectives, including those of different HCPs and patients, or that address the dynamics of HCP-patient relationship for designing XAI systems
c) Neglect of implementation and adoption challenges: Real-world barriers such as workflow integration, deployment feasibility, and post-deployment feedback mechanisms are often overlooked, despite their importance for clinical adoption.
d) Domain bias toward medical imaging: Prior reviews heavily focus on imaging-based applications, with limited synthesis of XAI use in non-imaging domains.
e) Insufficient evaluation of trust and usability: Few studies explore how trust and usability are assessed, or how faithfulness and plausibility are balanced to foster trustworthy AI.
f) Superficial coverage of human–AI interaction: Key elements such as user-centered design, explanation strategies, and cognitive factors influencing explanation effectiveness are underexplored.
g) Inconsistent use of core terminology: Terms like explainability, interpretability, and transparency are often used interchangeably, leading to conceptual ambiguity and lack of standardisation.

2.4. Research Aim and Objectives

Research Aim:

This SLR aims to systematically investigate the role and effectiveness of XAI methods in healthcare, focusing on their ability to support clinical decision-making, enhance diagnostic accuracy, enable personalised care, and meet the interpretability, usability, and trust needs of HCPs and patients across diverse healthcare domains, while emphasising the balance between faithfulness and plausibility in fostering appropriate trust.

Research Questions:

RQ 1: What evidence does the literature provide about the effectiveness of XAI methods in supporting clinical decision-making, improving diagnostic accuracy, and enabling personalised care for HCPs and patients in different healthcare domains?

RQ 1.1: Which healthcare domains and stakeholder groups have adopted XAI methods, and what are its key use cases in supporting clinical decision-making, diagnostic accuracy, and personalised care?

RQ 1.2: What evaluation metrics and methodologies are used to assess the technical performance, usability, and clinical impact of XAI models in healthcare, and how do these approaches support real-world applicability?

RQ 2: How do user studies evaluate whether XAI meets the interpretability and usability needs of both clinicians and patients?

RQ 2.1: What study designs, user-centred approaches, and experimental frameworks are used to assess the usability and interpretability of XAI in healthcare?

RQ 2.2: What challenges are faced by HCPs and patients for real-world adoption of XAI tools?

RQ 3: How do different qualitative and quantitative methods contribute to assessing different dimensions of trustworthiness?

3. Methodology

This SLR followed Kitchenham et al.'s [43] guidelines to summarise existing studies on XAI in healthcare, focusing on its role in clinical decision-making, diagnostic accuracy, and personalised care while emphasising the evaluation of transparency, interpretability, and reliability in AI systems for providing clear, understandable explanations of model decisions, fostering trustworthiness and user confidence. This is a methodological and usability-focused systematic review related to the healthcare domain. This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [44].

A SLR is a method of "identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest" [43] and includes the following phases 1. Identification of research 2. Study Selection 3. Quality Assessment 4. Data extraction 5. Data Synthesis 6. Threat to Validity

3.1. Identification of research

For the identification of relevant studies, this review has used an unbiased search strategy. The search utilised a combination of relevant keywords and their synonyms within this research domain. The search string was formulated using the following four distinct groups; Group 1 consists of keywords relevant to XAI and Related Concepts, Group 2 consists of keywords relevant to Healthcare and Related Domains, Group 3 consists of keywords relevant to

Evaluation Metrics and Challenges AND Group 4 User Studies & Usability in XAI. Boolean operators were employed to combine these search terms, resulting in the following search query as mentioned in Table 2. This comprehensive search strategy aimed to identify relevant studies on XAI in this area of research from well-known online databases; Web of Science, Scopus, IEEE Xplore, PubMed, Cochrane Library, and ACM Digital Library. This comprehensively covers the fields of computer science, healthcare, and artificial intelligence. To be inclusive and structured this review has used the same keywords on all six databases and a search was conducted within the title, abstract, and author keywords.

To streamline the systematic literature process two key software Rayyan[1] and Mendeley[2] were used. Mendeley, a reference management software, was used for managing references, detecting duplicates, and organising full-text articles during the full-text review process. Rayyan facilitated the initial screening of titles and abstracts, detected duplicates, and allowed researchers to annotate excluded studies for future reference.

Table 2: Query formation

Groups	Search Terms/Keywords
Group 1: XAI and Related Concepts	Explainable AI" OR "XAI" OR "Explainable Artificial Intelligence" OR "Interpretable AI" OR "Transparent AI" OR "AI explainability" OR "Model interpretability" OR "Interpretable machine learning" OR "AI transparency" OR "Transparent Decision Making" OR "Interpretable algorithms" OR "XAI frameworks" OR "Model transparency" OR "Explainability in AI
Group 2: Healthcare and Related Domains	"Healthcare" OR "Clinical decision-making" OR "Health technology" OR "Diagnostic accuracy" OR "Personalised care" OR "Personalised treatment" OR "Precision medicine" OR "Medical decision support" OR "Health informatics" OR "Medical AI applications" OR "Patient care" OR "Digital health" OR "Clinical

[1] https://www.rayyan.ai/

[2] https://www.mendeley.com/search/

	decision support system" OR "CDSS" OR "Clinical diagnosis" OR "Disease prediction" OR "Healthcare analytics" OR "Clinicians" OR "Healthcare providers" OR "Doctors" OR "Physicians" OR "Nurses" OR "Medical professionals" OR "Healthcare workers" OR "Health practitioners" OR "Specialists" OR "Surgeons" OR "General practitioners" OR "GPs" OR "Patients" OR "Patient populations" OR "Healthcare consumers" OR "End-users" OR "Medical staff" OR "Care providers" OR "Primary care physicians" OR "Medical teams" OR "Medical personnel" OR "Therapists" OR "Psychologists" OR "Health service users" OR "Health patients" OR "Pharmacists" OR "Health equity" OR "Health disparities" OR "Clinical trials" OR "Medical research" OR "Biomedical research" OR "Primary care" OR "Telemedicine" OR "Hospital settings"
Group 3: Evaluation Metrics and Challenges	"Evaluation metrics" OR "Model accuracy" OR "Interpretability" OR "Fidelity" OR "Trustworthiness" OR "Trust" OR "Comprehensibility" OR "Actionability" OR "Clinical validation" OR "Model performance" OR "Challenges" OR "Implementation barriers" OR "Black-box models" OR "Ethical AI" OR "Bias in AI" OR "Data privacy" OR "Regulatory issues" OR "Clinician trustworthiness" OR "Model complexity" OR "Explainability-accuracy trade-off" OR "Clinician-centric needs" OR "Patient-centric needs" OR "Performance metrics" OR "AI ethics" OR "Clinical effectiveness" OR "AI governance"
Group 4: User Studies & Usability in XAI	"User study" OR "Evaluation study" OR "User studies" OR "User evaluation" OR "Human-centred AI" OR "User-centred design" OR "User experience" OR "Cognitive load" OR "Usability testing" OR "Human factors" OR "Clinician feedback" OR "Patient feedback" OR "Heuristic evaluation" OR "Experimental study" OR "Qualitative evaluation" OR "Quantitative evaluation" OR "Think-aloud study" OR "Survey-based study" OR "Interview-based study" OR "Focus group" OR "Usability metrics" OR "Interaction design" OR "Human-AI interaction" OR "Explainability assessment" OR "XAI user experience" OR "Clinician XAI adoption" OR "Patient XAI adoption"

Query	(Group 1) AND (Group 2) AND (Group 3) AND (Group 4)

3.2. Study Selection

3.2.1 Study Selection Criteria

For the identification of primary studies well-defined inclusion and exclusion criteria based on the scope of this review work were outlined as follows:

Inclusion Criteria

- Peer-reviewed articles and conference papers focusing on XAI in healthcare.
- Studies published in the last 10 years (2014–2024) to capture recent advancements in XAI, particularly evaluating explanation performance and its impact on users.
- Studies published in English.
- Studies involving HCPs (e.g., clinicians, physicians) or patients and focusing on diverse healthcare settings (e.g., hospitals, clinics, digital health).
- Studies discussing specific use cases of XAI in healthcare, particularly related to clinical decision-making, diagnostic accuracy, personalised care, or disease prediction.

Exclusion Criteria

- Editorials, opinion pieces, blog posts, surveys, reviews, book series, letters, and commentaries that do not provide empirical contributions to the understanding of XAI in healthcare.
- Duplicate reports of the same study.
- Studies focussing solely on traditional AI models without emphasising the need for or use of XAI methods.
- Studies solely addressing non-healthcare applications of XAI or lacking a strong connection to the topic, despite containing relevant keywords in the initial search.
- Studies focusing on medical imaging and computer vision techniques due to a prevalent bias in the literature towards these applications.
- Studies involving non-human subjects.
- Studies explicitly comparing XAI models with traditional black-box models in terms of performance or applicability.

3.2.2 Study Selection Process

A total of 186 studies relevant to the research topic were identified in the initial search by applying a comprehensive search string (Table 2) to the titles, keywords, and abstracts of articles, while also applying filters for publication year (2014-2024) and language (only English). A systematic selection process, outlined in the PRISMA flow diagram (Fig. 1), ensured a rigorous and replicable review.

Duplicate Removal: Duplicates were identified and removed using automated tools such as Rayyan and Mendeley, resulting in 121 unique articles.

Title and Abstract Screening: Each title and abstract was evaluated against predefined inclusion and exclusion criteria, leading to the retrieval of 51 preliminary studies, which include journals, conferences, and book series. Subsequently, 43 studies were available for full-text review through open access or library access.

Full-Text Review: 21 studies were excluded since they focus solely on evaluating traditional model performances, rather than evaluating the quality of explanations, resulting in 22 final

studies: 15 peer-reviewed journals and 7 conference proceedings.

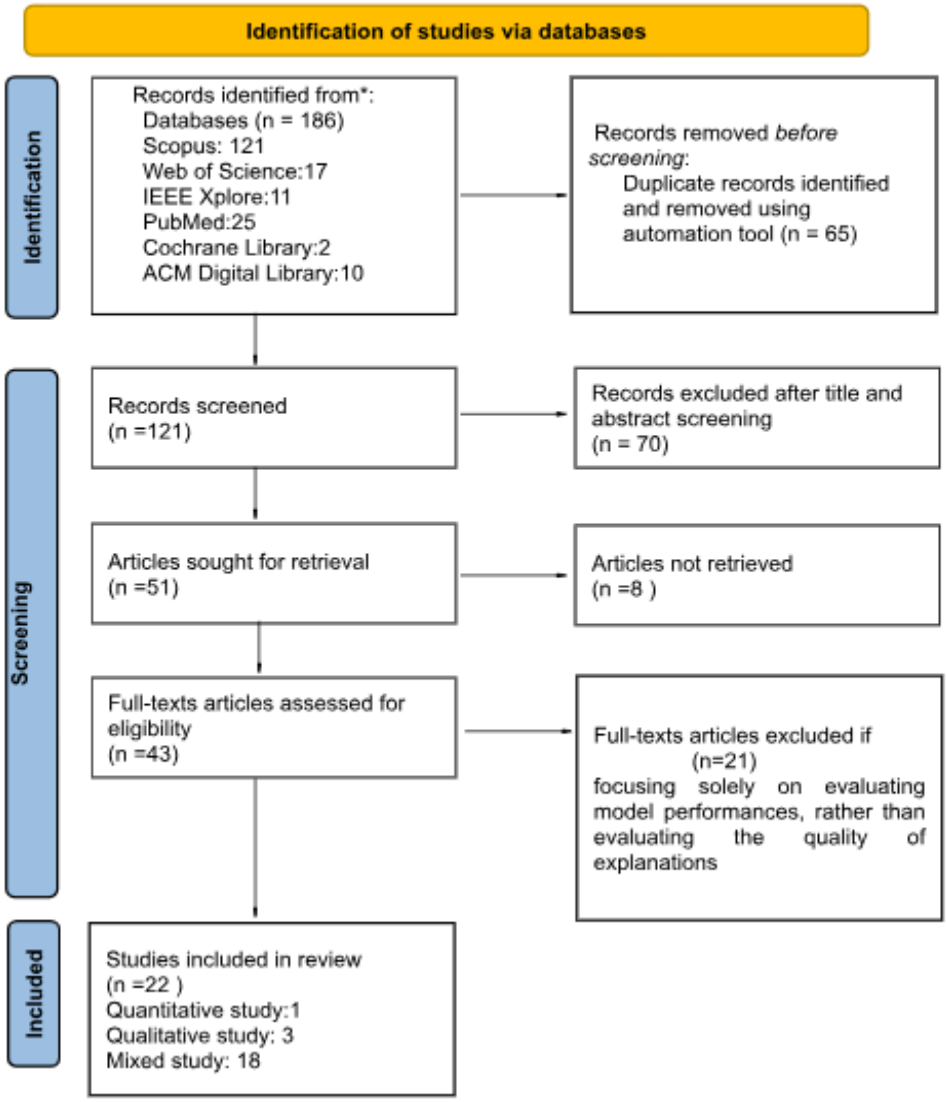


Fig 1: PRISMA Flow Diagram

3.3. Quality Assessment

In this phase, the quality of selected studies was assessed to help in the interpretation of findings and determine the strength of inferences. Although, there is no defined definition of study quality, both the CRD Guidelines [45] and the Cochrane Reviewers' Handbook [46] highlight that quality is determined by how effectively a study minimises bias while maximising internal and external validity. Hence, the quality of each included study was assessed using appropriate

quality appraisal tools based on the study design [47]. In this review, the study designs of the selected studies are categorised as quantitative study, qualitative study and mixed methods. Therefore, the Mixed Methods Appraisal Tool (MMAT) 2018 [48], a critical appraisal tool designed to assess the quality of qualitative, quantitative, and mixed-methods studies was used for a systematic evaluation of the evidence in terms of relevance, reliability, validity, and applicability. A consensus approach among the reviewers was used to evaluate study quality and mitigate bias. The results of this quality assessment were furthermore, considered in the synthesis process which ensured robust and reliable findings. It is important to note that MMAT discourages to exclude studies with low methodological quality; instead encourages to provide a more detailed presentation of the ratings of each criterion to better inform the quality of the included studies [48]. This was done using a Microsoft Excel spreadsheet, which facilitated the tracking and analysis of the assessment results. The list of the quality assessment questions was presented below for Quantitative Study, Qualitative study and Mixed methods study. Furthermore, the quality assessment was reproducible, ensuring consistent results upon repetition (see Supplementary Material for quality assessment of each study)

Quantitative Studies

QAQN1: Is the sampling strategy relevant to address the research question?
QAQN2: Is the sample representative of the target population?
QAQN3: Are the measurements appropriate?
QAQN4: Is the risk of nonresponse bias low?
QAQN5: Is statistical analysis appropriate to answer the research question?

Qualitative Studies

QAQL1. Is the qualitative approach appropriate to answer the research question?
QAQL2. Are the qualitative data collection methods adequate to address the research question?
QAQL3. Are the findings adequately derived from the data?
QAQL4. Is the interpretation of results sufficiently substantiated by data?
QAQL5. Is there coherence between qualitative data sources, collection, analysis, and interpretation?

Mixed Methods Studies

QAM1. Is there an adequate rationale for using a mixed methods design to address the research question?
QAM2. Are the different components of the study effectively integrated to answer the research question?
QAM3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

QAM4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed
QAM5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

The quality assessment matrix (Supplementary Material) indicates that most studies employed mixed methods study design and they consistently met high standards across all evaluation criteria (QAM1–QAM5). These studies demonstrated strong rationale (e.g., [50], [51], [55]), effective integration of qualitative and quantitative data (e.g., [57], [62], [70]), and a reflective interpretation of divergences (e.g., [63], [68]). For instance, [55] combined quantitative trust scores with physician feedback to assess XAI usability, while [70] used HCP and patient input alongside system performance metrics to evaluate a diabetes dashboard. However, a few studies showed limitations. For instance,[52],[55], and [59] were rated "No" on QAM4, indicating that they failed to adequately address divergences between qualitative and quantitative findings. [61] rating “Can’t tell” on QAM3, QAM4, and QAM5, suggest insufficient clarity in data integration and methodological adherence. These weaknesses, while few, highlight the importance of transparent interpretation and synthesis in mixed methods studies. For example, [55] assessed XAI usability with strong rationale and integration but didn’t fully explore inconsistencies between data sources. Overall, the high quality of most mixed methods studies supports the robustness of their insights into XAI in healthcare, though attention to integration gaps remains necessary. Qualitative studies, though fewer (e.g., [54], [64]), showed methodological rigor, with appropriate data collection methods and coherent interpretation whereas the quantitative study [53] exhibited limitations, such as unclear sampling strategies and potential nonresponse bias, which may affect generalisability. Overall, the methodological quality was strongest in mixed methods and qualitative studies, offering robust insights into the evaluation and usability of XAI in healthcare. This quality-based assessment emphasised on not relying heavily on weaker studies in the synthesis because of methodological shortcomings. However, the overall methodological quality of the included studies was high, with only a small proportion identified as lower quality ([52],[53],[55],[59],[61]). These studies were retained for completeness but contributed minimally to the synthesis. Hence, the main findings of the review are based on methodologically robust evidence.

3.4. Data extraction

The data was extracted for 22 papers and the formulation of the data extraction strategy was guided by the underlying research questions. This structured form was designed to capture key information necessary for achieving the objectives of the study. The data was systematically logged using a tabulated Microsoft Excel spreadsheet for efficient management and analysis. A pilot study was first conducted on a few preliminary studies by each of the reviewers to mitigate

bias. A unique identifier was assigned to each study that was made up of the initial letter of a source followed by a serial number. The form includes study identifiers, publication details, research questions, datasets, methodologies, algorithms used, evaluation metrics, findings, primary beneficiaries, study limitations, etc. The extracted data encompassed a variety of formats, including textual descriptions, URLs, and numerical values, allowing for structured organization and analysis. Furthermore, the quality assessment checklist is included within the data extraction form (Supplementary Material) to ensure that each study is evaluated for reliability and rigor, improving the overall quality of the review process [49]. Table 3 provides a detailed description of specific fields extracted from these studies with the consensus among all the authors.

Table 3. Elements of the study.

Elements	Description
Study ID	Unique identifier for the studies
Year of Publication	Year the study was published
Journal/Source	The journal or publication source where the study appeared, e.g., Nature, IEEE Access.
Type of publication	Type of publication, e.g., Journal article, Conference paper.
Database	Database used for retrieval, e.g., Scopus, Pubmed, IEEE.
Objective/Research Question	Main objective or research questions addressed
Dataset	Dataset used in the study
Source code	Availability of source code related to the study
Study design	Type of study (e.g., experimental, observational, qualitative, etc.)
Population/Participants	Description of the study population, including sample size
Data Collection/ Sampling Methods	Techniques used for data collection (e.g., survey, experiment, etc.)
Data Analysis Methods	Statistical or analytical methods used
Algorithms used	Which XAI algorithms were used?

Application area	In which medical domain does the study intend to be applicable
Model performance metrics	Evaluate model performances like Accuracy, Sensitivity, etc.
XAI Performance Evaluation metrics Objective (Computer-based Evaluation Metrics): Model-based, Attribution-based, Example-based Subjective (Human-Centred Evaluation Metrics): Explanation Evaluation, Usability Evaluation	Evaluates the quality of explanations
Quantitative methods	Quantitative methods used for measuring different explainability metrics
Qualitative methods	Qualitative methods used for measuring different explainability metrics
XAI Properties	Interpretability, Transparency, Trustworthiness, Fairness, Accountability, Causality, Actionability etc.
Key Findings	Summary of primary findings
Limitations	Limitations addressed by the study
Consistency with Literature	Whether findings are consistent with existing literature
Potential Conflicts of Interest	Disclosed conflicts of interest
Primary Beneficiary	Key groups or stakeholders who have benefited from the study's findings or applications.
Quality Assessment QAQN1-QAQN5 (Quantitative) QAQL1-QAQL5 (Qualitative) QAM1-QAM5 (Mixed Methods)	Questions derived from Mixed Methods Appraisal Tool (MMAT) 2018 [50], a critical appraisal tool (Yes/No/Can’t tell)
Relevance to SLR	Yes/No - whether the study meets the SLRs main focus
Link	Link to the study

3.6. Threat to Validity

This SLR aims to provide a comprehensive and unbiased synthesis of existing research on XAI in healthcare but several potential threats to validity must be acknowledged. Design and execution of the review process, as well as inherent limitations within the primary studies themselves led to these limitations.

Search String: While the search string employed in this review was designed to be as exhaustive as possible, it is possible that some synonyms were overlooked. This omission may have resulted in the exclusion of valuable studies relevant to our research domain.

Language Barrier: The inclusion criteria stipulated that only articles published in English were considered for this study. This language filter may have led to the omission of significant research published in other languages, which could have provided valuable insights into our research topic.

Time Frame of Selected Studies: This review only includes studies published in the last 10 years (2014-2024) to make this as comprehensive as possible. However, few studies published in 2025 might be relevant for this research topic which can consequently introduce a bias.

Selection and Publication Bias: The criteria for study inclusion and exclusion, along with reliance on reputable databases such as Web of Science, Scopus, IEEE, may introduce selection bias. This could result in a skewed representation of the available evidence. Furthermore, the exclusion of grey literature (e.g., editorials, opinion pieces, surveys, and reviews) may also contribute to publication bias, as valuable insights from non-peer-reviewed sources are omitted.

Domain-Specific Exclusion: Studies related to medical imaging and computer vision techniques have been excluded in this review because of extensive coverage of these topics in prior reviews. However, this exclusion may limit the comprehensiveness of the review with respect to the broader landscape of AI applications in healthcare.

Narrow Stakeholder Focus: It focuses only on two stakeholders in the healthcare domain; HCPs and patients but limiting the perspective of other stakeholders; healthcare administrators, developers, regulatory entities etc. This narrow scope may affect the generalisability of the findings to the wider healthcare ecosystem.

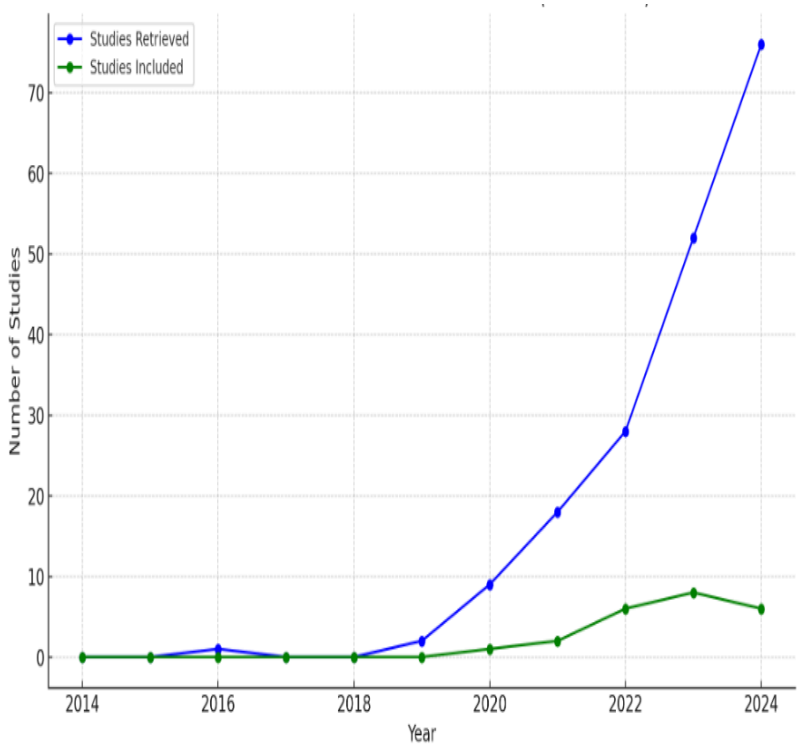
3. Results

3.1 Research Overview and Synthesis

This section presents a comprehensive overview of current research on XAI in healthcare by organising, summarising, and integrating findings from the studies . The aim is to provide a clear understanding of research trends, study characteristics, and key themes within this emerging domain.

3.1.1 Research Trends

The research trend in this domain has shown a notable increase since 2020, with a significant surge in publications since 2023 (Fig. 2a). Table 4 summarises the distribution of retrieved and included studies across databases, with Scopus providing the largest number of relevant articles, followed by ACM Digital Library and IEEE Xplore. This spread highlights the multidisciplinary nature of research spanning engineering, biomedical, and clinical domains.The studies cover a broad range of healthcare domains (Table 5), including Intensive Care Medicine, Paediatrics, Neurology, Oncology, Cardiology, and Clinical Decision Support Systems (CDSS), as well as interdisciplinary topics such as Human-AI collaboration and Human-Centred AI (HCAI). Study designs predominantly use mixed methods (81.8%), with smaller proportions of qualitative (13.6%) and quantitative studies (4.5%) (Fig. 2b). Overall, these patterns suggest an emerging research field that not only demonstrates increasing volume but also domain diversity, particularly in the years 2023 and 2024.


Studies Retrieved
Studies Included
Number of Studies
Year
0
10
20
30
40
50
60
70
2014
2016
2018
2020
2022
2024

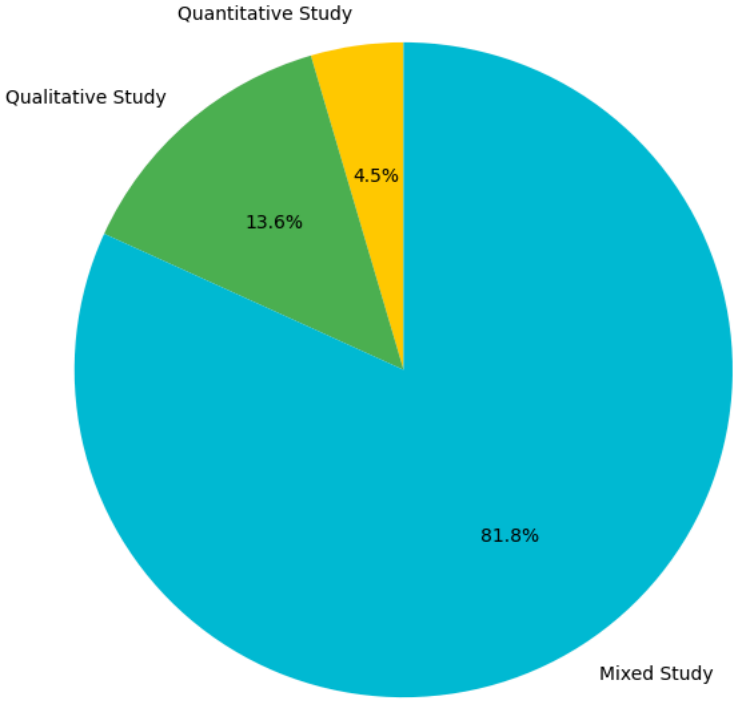

Quantitative Study
Qualitative Study
4.5%
13.6%
81.8%
Mixed Study

Fig. 2. (a) Trend of retrieved and included studies (2014–2024) (b) Distribution of study designs of included studies (2021-2024)

Table 4. Summary of search results (2014-2024) and includes studies (2021-2024)

Database Engines	Source Address	Number of search results	Number of relevant articles
Scopus	https://www.scopus.com/	121	11
Web of Science	https://www.webofscience.com/	17	1
Pubmed	https://pubmed.ncbi.nlm.nih.gov/	25	2
Cochrane library	https://www.cochranelibrary.com/	2	0
ACM Digital Library	https://www.acm.org	10	6
IEEE Xplore	https://ieeexplore.ieee.org	11	3

Table 5: Research domains in healthcare in each year (2021-2024)

Year of Publication	Healthcare Domain	Subdomains	Study	Conference proceedings	Peer-reviewed journals
	Intensive Care	Sepsis	[51]	0	1

Year of Publication	Healthcare Domain	Subdomains	Study	Conference proceedings	Peer-reviewed journals
2024	Medicine (ICU)				
	Paediatrics	Neonatology	[52]	1	0
	Neurology	Cognitive Disorders, particularly Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD)	[56]	1	0
	Nephrology	Polycystic Kidney Disease (PKD), particularly Autosomal Dominant Polycystic Kidney Disease (ADPKD)	[59]	0	1
	Infectious Diseases	Hospital-onset Bacteraemia (HOB)	[61]	0	1
	Healthcare Analytics	Predictive Health Technologies with a focus on XAI for CDSS.	[69]	1	0
				3	**3**
2023	Endocrinology	Diabetes Care	[50]	0	1
	Oncology	Lung Cancer	[54]	0	1
	Medicine	CDSS	[57]	0	1
	Cardiology	Electrocardiography (ECG), ECG (electrocardiogram) signal classification and cardiac abnormalities.	[60],[67]	1	1
	Neurology	Epilepsy and Seizure Detection	[65]	0	1
	CDSS+ Healthcare Analytics	Explainable AI(XAI) applied to healthcare decision-making, Predictive Health Technologies with a focus on Diabetes Risk Prediction and Explainable AI(XAI) in CDSSs	[68],[70]	1	1
				2	**6**
2022	Neurology	Amyotrophic Lateral Sclerosis (ALS).	[53]	1	0
	General Medicine	Electronic Health Records (EHRs) and CDSSs.	[55]	0	1
	Obstetrics	Gestational Diabetes Mellitus (GDM).	[62]	0	1
	CDSS	Healthcare Providers' Decision-Making, ML in Healthcare and Electronic Health Records (EHR)	[63],[64]	0	2
	Ophthalmology	Age-Related Macular Degeneration (AMD)	[66]	1	0

Year of Publication	Healthcare Domain	Subdomains	Study	Conference proceedings	Peer-reviewed journals
				2	**4**
2021	Medicine	Paediatrics	[58]	0	1
	Neurology	CDSSs with a focus on Explainable AI(XAI) and Decision-Making Support for Neurologists.	[71]	0	1
				0	**2**

3.1.2 Evaluation Approaches

Studies evaluate AI systems in healthcare using model performance metrics to assess predictive accuracy and XAI evaluation performance metrics to assess the effectiveness of explainability. Model performance metrics like accuracy [50,52,54,56,59,60,66,71], AUROC [50,51,55] etc., are used to evaluate the predictive capabilities of AI models, and XAI evaluation performance metrics are used to assess the quality of explanations. The classification of XAI evaluation metrics into objective and subjective (or human-centered) categories allows for a comprehensive assessment of both computational performance and complexity of the explanation model, as well as how well the users comprehend these explanations. Objective metrics focus on the technical attributes of the explanation, such as fidelity to the underlying model, complexity, stability, and computational efficiency [e.g., 67,68]. These are typically assessed through computer-based, quantitative approaches. In contrast, subjective or human-centered metrics evaluate how end-users, such as clinicians and patients, perceive, interpret, and interact with the explanations. These include aspects like interpretability, trust, satisfaction, cognitive workload, and perceived usefulness, predictability, consistency, ecological validity and practicability [e.g.,55,57]. Although subjective in nature, these metrics are often measured quantitatively through methods such as Likert-scale surveys, task performance scores, and usability questionnaires. Additionally, qualitative methods such as interviews, think-aloud studies, and open-ended feedback are used to gain deeper insights into user experiences and expectations (See Appendix Table A.1 for more details).

In this review , subjective metrics have been categorised into explanation evaluation metrics and usability metrics, which address distinct yet interrelated aspects of AI system performance. Explanation evaluation metrics ensure that the AI's reasoning is clear, trustworthy, and helpful, which is vital for user confidence whereas usability metrics focus on how intuitive, efficient, and practical the system is, ensuring that clinicians can easily integrate it into their workflow. This classification will help the researchers to understand the gaps in specific aspects of subjective

metrics, ensuring that AI systems are both interpretable and user-friendly for effective real-world application since a system with strong explanations but poor usability may be difficult to use, while a highly usable system with unclear explanations may lead to distrust. Fig 3, shows this classification in a systematic way to understand the illustration for AI evaluation. This review has furthermore, identified several objective and subjective metrics and Fig 4 shows the mapping of studies with broader classification of metrics which unveils that objective evaluation metrics have been less used in these studies than subjective metrics, focusing more on comprehending the explanations from a human perspective.

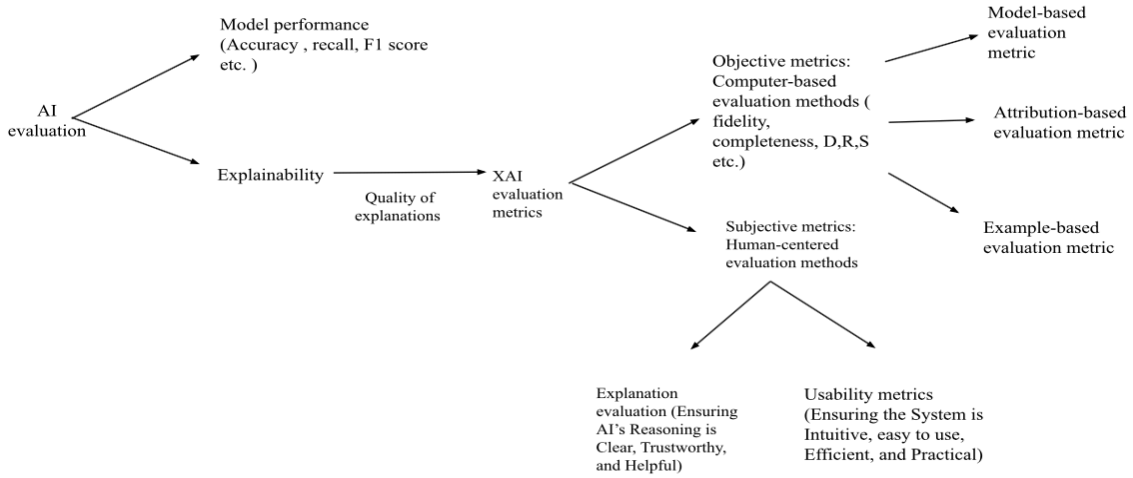


Fig 3: Illustration for AI Evaluation

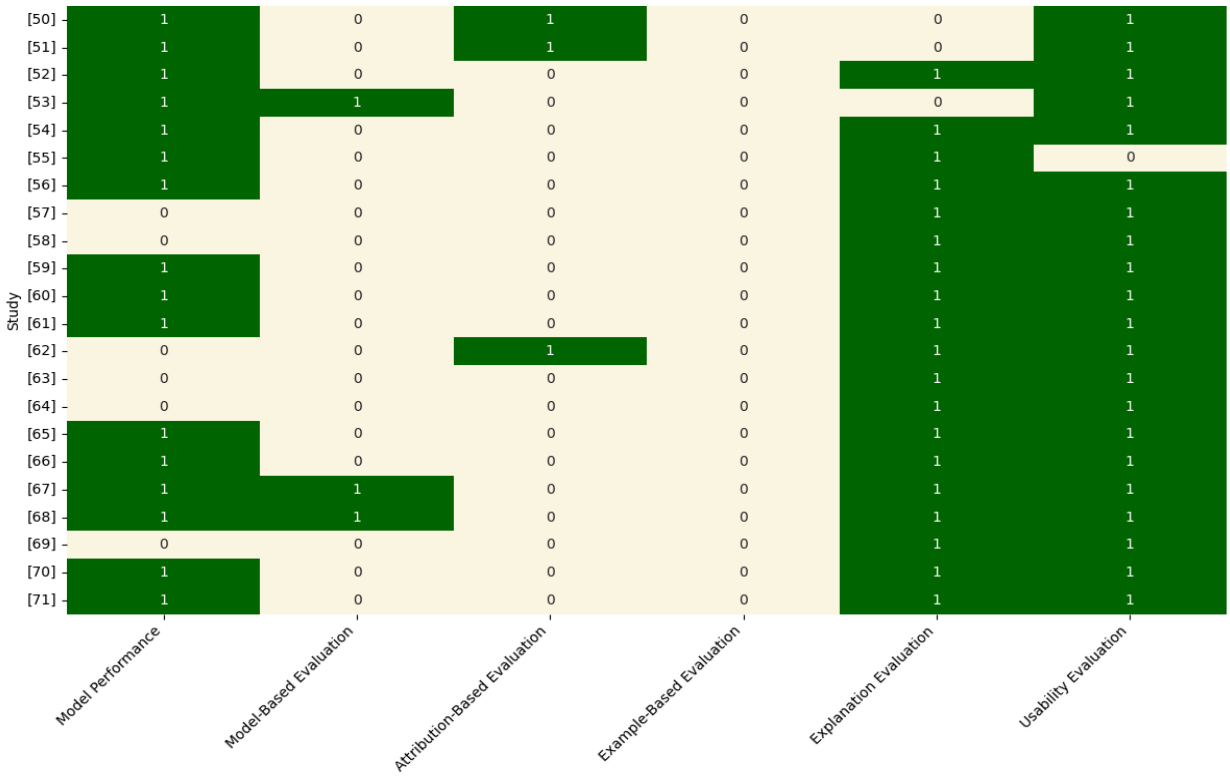


Fig 4: Study-Metric Adjacency Matrix

3.1.3 XAI Properties

XAI properties, including transparency, interpretability, trustworthiness, and actionability, are crucial in understanding how explanations influence HCPs decision-making. The studies are skewed towards trustworthiness, transparency, and interpretability, emphasising improving users' trustworthiness, decision-making, and understanding of AI-based healthcare systems. Explainability, usability, causality, and usefulness are also crucial XAI properties, while properties like actionability, faithfulness, personalisation, and clarity receive moderate attention. Fairness, justifiability, adaptability, confidence, interactivity, comprehensibility, and flexibility are explored to a much lesser extent (Table 6) in this area of research. Overall, these patterns indicate that while foundational XAI properties receive significant attention, aspects related to user experience and rigorous explanation validation remain underexplored, highlighting important directions for future research.

Table 6: Overview of Studies Addressing Key XAI Properties

XAI Property	Studies
Trustworthiness	[50], [52], [55], [56], [57], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [70], [71]
Transparency	[50], [54], [55], [56], [57], [58], [61], [62], [63], [64], [66], [68], [69]
Interpretability	[50], [53], [54], [55], [56], [59], [61], [62], [63], [64], [65], [68], [69]
Explainability	[53], [55], [58], [61], [62], [66], [71], [68]
Usability	[50], [53], [60], [64], [57]
Causality	[52], [53], [59], [64], [71]
Usefulness	[51], [52], [67], [70], [63]
Actionability	[51], [52], [67], [70], [63]
Faithfulness	[51], [55], [68]
Personalisation	[56], [62], [71]
Clarity	[62], [66], [67]
Justifiability	[62], [63]
Fairness	[52], [61]
Adaptability	[54], [58]
Confidence	[52], [65]
Interactivity	[56], [66]
Comprehensibility	[60]
Flexibility	[55]

3.1.4 Stakeholders and Beneficiaries

The primary beneficiaries in these studies are grouped in Table 7. It suggests that healthcare providers such as clinicians, medical specialists (e.g., neurologists, oncologists, neonatologists), nurses, and EMTs are dominant beneficiaries who benefit from AI systems that support diagnostic and treatment decisions, particularly through interpretable outputs that align with clinical reasoning. Non-healthcare stakeholders (AI Researchers, Policymakers, Developers) ([52], [53], [54]) focus on technical development, policy formation, and the integration of XAI into healthcare infrastructure. Patients are explicitly identified as beneficiaries in only two

studies ([67], [70]), highlighting a significant gap in patient-centered AI design. While healthcare providers remain the primary users, the limited inclusion of patients suggests that it is important to include them as active stakeholders, as AI becomes increasingly involved in personal health management and decision-making.

Table 7: Stakeholder Groups in Healthcare and AI Integration

Stakeholder groups	Stakeholder	Study	Primary Beneficiary
Healthcare	Healthcare Providers, Clinicians, & Medical Professionals	[50], [51], [53],[54],[55],[56], [57], [58],[59],[60], [61],[62], [63], [64], [65], [66], [67],[68], [69], [71]	Medical professionals in neonatal care , oncology community, and clinicians (particularly oncologists), physicians and nurses, HCPs, especially neurologists, medical professionals, specifically those involved in clinical decision-making, clinicians using AI-based decision support systems, particularly in child health, doctors and healthcare providers, especially in nephrology, ECG readers (novices and experts), health institutions, clinicians & healthcare providers for HOB risk prediction, healthcare practitioners (obstetricians, dietitians, and other medical professionals) involved in the care of pregnant women, especially those at risk for gestational diabetes mellitus (GDM), healthcare assistants, medical practitioners (doctors, nurses, EMTs).
Non-Healthcare	AI Researchers, healthcare Policymakers, Developers	[52],[53], [54]	AI researchers and policymakers focusing on AI applications in healthcare, including oncology and clinical decision-making, developers
Patients	Patients	[67],[70]	Patients benefiting from more transparent AI-driven ECG diagnosis, Patients at risk for diabetes

3.1.5 Study Limitations

The studies reviewed also highlight a considerable number of limitations within this research area, which have been categorised and summarised in Table 8, with Figure 5 showing the number of studies per year reporting at least one of these limitations noted in Table 8. These limitation categories illustrate a multi-faceted landscape of challenges ranging from methodological rigor and data quality to user interaction and scalability that need to be addressed to advance the reliability, usability, and clinical integration of XAI in healthcare. For instance, many studies struggled with the quality and clarity of AI explanations, which can hinder user trust. limitations in real-world applicability also highlight the gap between controlled research settings and clinical practice, etc.. Therefore, the focus is moving toward solving real-world problems and improving study quality, but there are still gaps in implementation and reproducibility planning.

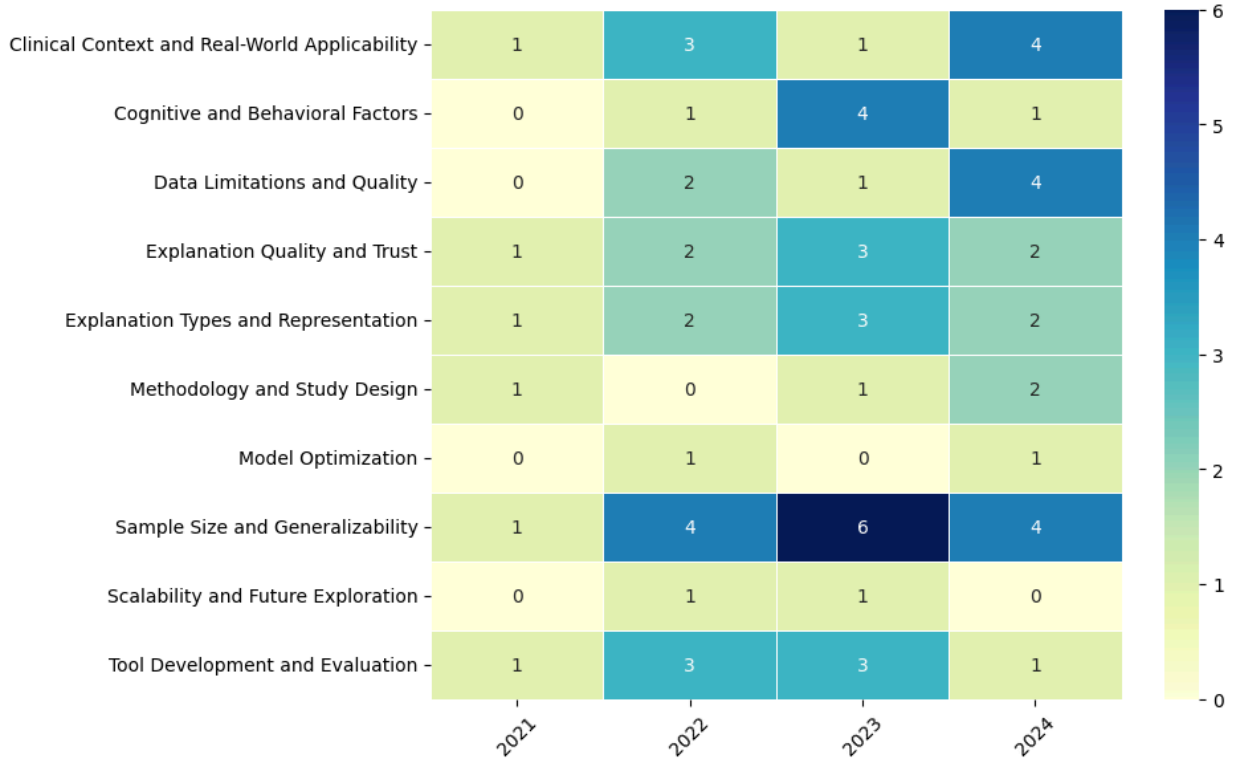


Fig. 5 Visualises the frequency of limitations reported annually.

Table 8: Key Limitations in AI and Healthcare Research

Limitation Category	Description	Study
Sample Size and Generalizability	Limitations related to the number of participants in a study and how well the study's findings can be applied to larger or different populations.	[53], [58], [59], [60], [61], [62], [63], [64], [65], [66], [68], [69], [70], ,
Methodology and Study Design	Limitations related to the design and methodology used in the study like issues such as bias introduced by how the study was set up,potential flaws in data collection methods, lack of control over confounding variables and inaccurately measuring or evaluating the outcomes of interest due to design issues.	[54], [57], [59], [61]
Explanation Quality and Trustworthiness	Limitations related to the quality, clarity, and effectiveness of the explanations provided by AI or decision support systems. It addresses how well the explanations help users understand the AI system’s decision-making process.	[53], [55], [58], [59], [61], [66], [71],
Clinical Context and Real-World Applicability	Limitations related to how well the study’s findings apply to real-world clinical settings. It includes concerns related to the applicability of the findings across different clinical contexts, where real-world conditions, workflows, and constraints may not align with the study setup, limiting the broader applicability of the findings.	[54], [55], [56], [60], [64], [65], [66],
Data Limitations and Quality	Limitation related to missing data, incomplete datasets, concerns about the diversity and quality of	[54], [55], [56], [62], [67], [68],

	the data, or the presence of errors in the data (e.g., misclassifications or artifacts).	
Tool Development and Evaluation	Limitations related to the development, testing, and evaluation of the tools or systems being studied. It includes concerns about the immaturity or inadequacy of the tools, absence of a common framework for evaluating tools or systems in the relevant field, the limited scope of their evaluation, or the lack of comprehensive comparison with other existing methods.	[53], [54], [58], [65], [69], [71],
Cognitive and Behavioural Factors	Limitations related to human cognition and behaviour, particularly in how individuals interact with the system. It considers issues such as cognitive biases, over-reliance on AI, and participants' ability to introspect and report their thought processes accurately.	[55], [57], [60], [63], [67], [71]
Scalability and Future Exploration	Limitations related to the scalability of the methods, tools, or systems used in the study. It includes concerns about whether the tools can be effectively scaled to larger datasets or more diverse real-world environments and also includes challenges such as expanding the methods to other domains, improving performance in more complex or varied settings, and exploring new avenues for tool or method optimization.	[67], [70]

Model Optimization	Limitations related to the optimization of the ML models used in the study. This includes a lack of fine-tuning, or the use of overly simplistic approaches that fail to capture the complexity of the data. It also involves concerns about how the models handle specific variables, temporal data, or rare events.	[54], [56]
Explanation Types and Representation	Limitations around the types and formats of explanations provided by AI systems. It covers whether the study only tested one form of explanation (e.g., textual explanations) and whether other types, such as visual or interactive explanations, were explored.	[53], [54], [63], [67], [68], [69],

3.1.6 Key Themes

This review identified several key themes across various studies and has revealed significant insights into its potential benefits and challenges due to integration of XAI into healthcare.

a) Trustworthiness and Confidence in AI Systems

- Increased Trustworthiness: Explanations play a crucial role in fostering trustworthiness in AI. Several studies show a clear pattern highlighting that the use of XAI leads to increased trustworthiness among users [50,60,68]. While HCPs [50] using the XAI4Diabetes app found predictions to be reasonable (90%), helpful, and easy to use (80%), [60, 71] highlighted that novice clinicians and cardiologists showed a significant increase in trustworthiness with more XAI tools compared to clinical experts. Trustworthiness was enhanced when explanations were understandable and AI predictions were easily interpretable for instance, the Global Tornado Plot made predictions easier to understand, enhancing trustworthiness [56, 60].
- Over-reliance on AI: While explanations can increase trustworthiness, they also raise concerns about over-reliance on AI, particularly with certain types of explanations like counterfactuals and example-based methods [54,62]. Participants showed signs of automation bias where they would agree with AI recommendations even when those

might be incorrect indicating over-reliance on AI. Moreover, participants did not adjust their trustworthiness or decision-making based on the accuracy of the predictions due to lack of significant difference in advice-taking behaviour for both correct and incorrect predictions [62,68].

- Scepticism from Experts: Medical experts exhibited more scepticism towards AI predictions and explanations than the general population, often questioning the reliability and clinical applicability of the tools. [52, 56, 71] highlighted that experts were more critical of AI systems, particularly in terms of performance and classification accuracy. They often challenged AI predictions, especially when the explanations were unclear, inconsistent, or failed to match their medical expertise. This scepticism was accompanied by a preference for greater transparency in how AI predictions were made, with many experts emphasising the need for AI tools that complemented, rather than replaced, human reasoning.

b) Usability and User Preferences

- User-Friendly Design: Usability is critical for the adoption of AI in healthcare. Ease of use and system customization were key factors influencing the adoption and effectiveness of XAI tools in healthcare. [50,61] highlighted that XAI systems were generally easy to use, with participants highlighting their intuitive design and interactive features. Visual tools, such as the Global Tornado Plot, and interactive elements like decision path visualisations, significantly enhanced user engagement and trustworthiness in the AI's predictions [56, 61].
- Customised Explanations: The need for customisation emerged as a recurring theme across multiple studies [58,71], highlighting that different users required tailored explanations. For instance, while novices preferred simplified, example-based explanations that were easier to understand, experts needed more complex models with dynamic explanations and detailed insights on prediction accuracy. This variation in user needs signifies the importance of adaptable interfaces that can be customized to fit the user's level of expertise, ensuring that XAI tools are accessible and useful for different HCPs.
- Interaction with Explanations: The introduction of interactive features like counterfactuals, visual aids (e.g., feature importance, decision paths), and data-centric explanations (e.g., trend visualisations) significantly improved clinician engagement with the tool [61,67,69,70] enhancing usability and trustworthiness. In [70], counterfactual explanations (VC4) were found to be particularly useful, while personalised visual summaries (VC2) were preferred due to their clear layout and transparency. Similarly, [67] emphasises the effectiveness of data-centric explanations, noting that ClusteredSHAP not only improved computational efficiency by reducing processing time

by 55-56% , but also helped clinicians engage more effectively with the model's decisions. Users rated ClusteredSHAP positively in areas like accuracy, understandability, and trustworthiness. These findings suggest that incorporating interactive features can improve clinician interaction with AI, leading to better interpretation and adoption of model insights in medical decision-making.

c) Explanation Methods and Their Effectiveness

- Preference for Certain Explanation Types: Different users preferred different types of explanations. For instance, clinicians often preferred example-based explanations, case-based reasoning, or feature importance over complex counterfactuals or model-centric explanations [69, 61,62]. Simpler explanations, especially visual ones like decision paths [61], were often considered more useful for decision-making.
- Improvement with Model Training: Some studies observed that the effectiveness of explanations improved when the predictive models were trained over time. This shows better alignment with existing clinical knowledge. For instance, [51] found that the usefulness of the XAI explanations increased, with concept scores such as temperature and leukocytes aligning more closely with existing clinical knowledge as the model was trained over time. This was highlighted with a significant improvement in the usefulness score, enhancing accuracy of feature importance values.
- Challenges with Complex Models: Some studies highlighted that complex AI models (e.g., those using SHAP or LIME) may present challenges for understanding, especially for experts who might find these explanations less clear or inconsistent with their medical knowledge. In [52] experts criticised the AI's explanations for being unclear and inconsistent with clinical reasoning, especially when misclassifying data. [53] emphasised that while SHAP and LIME are useful, their complexity requires deeper analysis and input from clinical experts to ensure practical relevance. [54] showed that oncologists needed more transparency and clear explanations about AI predictions, as the existing models did not provide sufficient insights into how decisions were made. These challenges underscore the need for clearer, more understandable explanations in AI models used in healthcare.

d) Impact on Clinical Decision-Making

- Influence on Decision-Making: XAI tools strongly influenced HCPs decision-making process as evidenced in the studies. The more explainable the AI tool, the more likely healthcare providers were to take its advice into account in clinical decisions. For instance, [62] highlighted that XAI methods, such as feature contribution and

example-based explanations, influenced healthcare practitioners' decisions, with no significant difference in advice-taking between the two. Similarly, [64] highlighted that the integration of XAI into clinical workflows helped clinicians interpret and trustworthiness ML predictions, particularly in predicting surgical complications, impacting clinicians decision-making process.

- Clinical Utility vs. Research Potential: Several studies highlighted that even if AI systems have potential to help with clinical decisions, HCPs contradict the idea of using them in real world clinical practice, especially if they lacked sufficient accuracy or real-world applicability. Instead, HCPs saw these tools as more suitable for research settings, where experts can test and refine them before integration into clinical workflows that directly impact patient outcomes [52,54].
- Uncertainty and Predictive Accuracy: A recurring theme across studies was the need for greater transparency in AI predictions, particularly regarding the uncertainty of those predictions. Clinicians expressed a desire for inclusion of confidence intervals or some indication of uncertainty alongside with the predictions for more reliable AI recommendations. For instance, [54] used Think Aloud Protocol (TAP) to understand how oncologists' perceive about a lung cancer relapse prediction system, revealing that credibility and utility were key factors in their willingness to integrate AI into their workflow. This highlighted the importance of knowing the uncertainty behind their predictions through confidence intervals helped clinicians assess reliability. Similarly, [71] demonstrated that perceived explainability influenced clinicians' interactions with AI-based CDSS. While certain XAI techniques, like decision trees, were seen as more interpretable, explainability alone did not always improve performance, particularly when explanations lacked clarity or alignment with clinical reasoning. This study also emphasised that a one-size-fits-all approach to AI explanations does not work, as different users interpret and trustworthiness AI outputs differently. Hence, without this transparency, the tools were seen as incomplete, limiting their trustworthiness and utility in clinical settings.

e) Trustworthiness-Building through Transparency

- Need for Transparency: Studies exhibit that clinicians prefer transparency in AI systems. They prefer systems which clearly explain how predictions were made, including confidence intervals of the predictions and potential risks associated with them. This is essential for building trustworthiness in AI tools [54, 61, 71]. Oncologists [57] expressed the need for prediction accuracy, including confidence intervals, to enhance trustworthiness. Clinicians [61] highlighted transparency in risk factors for planning tailored interventions. Additionally, [71] found that trustworthiness was more closely

linked to understanding the underlying data rather than the complexity of the AI model itself.

- Trustworthiness in AI vs. Human Expertise: Trustworthiness in AI was strongly tied to how well the AI system complemented human expertise. Many clinicians preferred systems that included both AI predictions and explanations alongside human clinical judgment, especially when the AI system was transparent and well-explained [56, 61,71]. The studies highlight that trustworthiness in AI among medical professionals (like neurologists) is closely linked to expert oversight and the alignment of AI with human expertise like their clinical judgement [56,61] . Likewise, medical professionals in [71] showed scepticism toward AI, preferring systems that allowed for collaboration between AI and human decision-making.

f) Domain-Specific Challenges and Needs

- Domain Expertise and Trustworthiness: HCPs with extensive domain knowledge often found AI explanations lacking because it did not always align with their clinical experience [52, 60, 71]. Hence, this highlights the importance of involving clinical experts in the development and evaluation of AI systems to ensure they are clinically relevant and trustworthy.
- Personalised Predictions: In fields like oncology, where dynamic and patient-specific predictions are crucial, AI systems struggled to provide accurate personalised predictions [54, 65, 71]. For AI to be more effective, it must incorporate individual patient data and adapt to the evolving conditions of the patient.

g) Incorrect predictions and misleading explanations

- Misclassification and Errors: Several studies noted that AI models were prone to errors, such as misclassifying data or producing incorrect predictions, which highly impacts trustworthiness and clinical utility. These errors led to concerns about the system's ability to generalise across different patient scenarios [52, 53].
- Misleading Explanations: In some cases, explanations were deemed unclear, misleading, or irrelevant. This was especially true when AI systems highlighted irrelevant features or failed to explain their decisions in a way that made sense to the clinicians' domain knowledge [52, 53] leading to confusion and reduced trustworthiness. For instance, even if visual explanations were often seen as helpful, their rationale and connection to the classification outcomes highlighted irrelevant areas, like padding, which increased confusion [52].

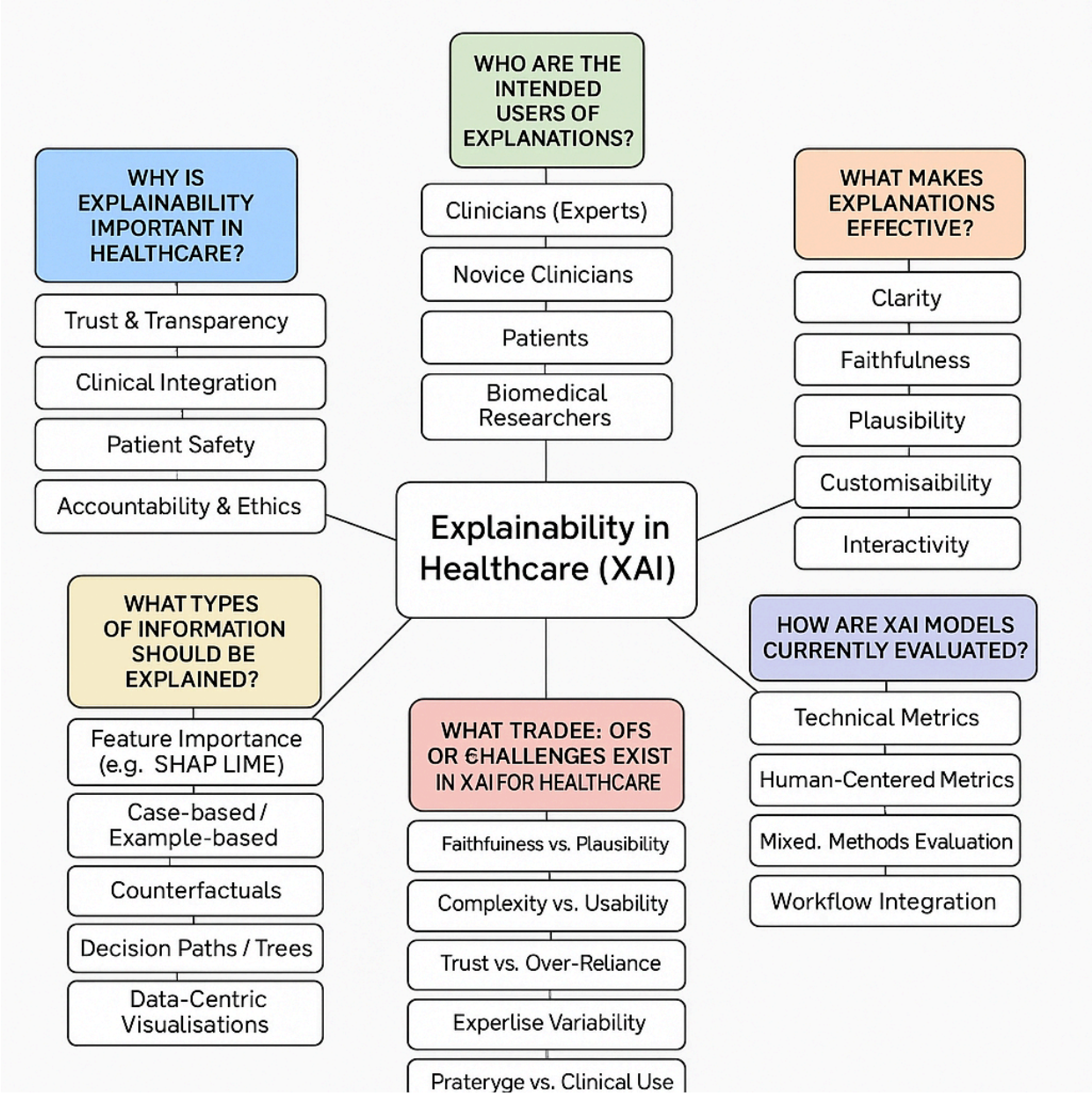
h) Perceptions of AI in Clinical Settings

- Differing Perceptions: Different healthcare stakeholders have different perceptions of AI in clinical settings. For instance, the general public and novice clinicians often responded positively to AI tools, perceiving them as useful and trustworthy, whereas expert clinicians were more critical, often questioning their reliability and clinical applicability [71, 60]. They evidenced that novice cardiologists had a significantly higher trustworthiness in AI compared to experts. Experts, while acknowledging AI's potential, often emphasised the importance of expert oversight and remained neutral to moderately positive in their trustworthiness.
- AI as a Tool for Support: There was a consistent perception that AI should complement human decision-making rather than replace it. This perspective was especially evident among medical professionals who emphasised the need for AI systems to remain under human oversight and guidance [60, 52, 71]. Studies such as [52,60] emphasise that clinicians see AI as a valuable support system that can augment their expertise, but they stress the need for human oversight, particularly when AI systems make errors. Experts prefer using AI as a decision support tool to assist with predictions, rather than fully trusting it to make critical decisions without human intervention. [71] indicates the necessity for customisation in AI systems, with medical professionals favouring methods like decision trees, which are perceived as more transparent, compared to more complex AI explanations like counterfactuals or case-based reasoning. Hence, customisable and transparent AI tools can ensure that AI tools can be aligned with human expertise for effective decision making.

i) Design Recommendations

- Iterative Design Process: [68] highlights the importance of iterative design processes which is important for improving the usability and effectiveness of AI explanation interfaces in CDSSs. This design process increases trustworthiness, addresses user needs, and identifies new technical requirements by prototyping, testing, and redesigning based on user feedback. This process helps refine both the user interface and the AI system, ensuring they align better with user expectations and real-world needs.
- Visualisation preference: In facilitating clinician engagement, trustworthiness, and informed decision-making clinicians favoured visual explanations, such as decision paths, global feature importance plots, and temporal views, over text-based explanations. These visualisations helped clinicians align their knowledge and clinical workflows with AI predictions, making them more intuitive and easier to interpret. For instance, Global Tornado Plot for neurologists [56] and reference-values for surgical predictions [64] help clinicians interpret AI outputs effectively, improving decision-making. Moreover, [70]

underscores the preference for clear visual layouts, like the VC2 patient summary and [61] highlights the importance of decision path visualisations for building trustworthiness in AI systems, as they make predictions easier to interpret.



3.2. Effectiveness of XAI in healthcare

This section presents our findings on the effectiveness of XAI methods in supporting clinical decision-making, improving diagnostic accuracy, and enabling personalised care for HCPs and patients in different healthcare domains. More specifically, we seek to answer RQ 2 using two sub questions: (1) which healthcare domains and stakeholder groups have adopted or explored

XAI methods and what are the key use cases (RQ1.1) and (2) what evaluation metrics and methodologies have been employed to assess the technical performance, usability, and clinical impact of XAI models (RQ1.2).

Supporting clinical decision making refers to systems designed to help HCPs (e.g., doctors and nurses) in making accurate and timely decisions regarding patient care. Increasing diagnostic precision refers to the ability of a system to improve the accuracy of identifying medical conditions or diseases. Enabling personalised care focuses on tailoring healthcare treatments and strategies to the unique characteristics, needs, and preferences of patients. The effectiveness of XAI methods in healthcare requires a nuanced understanding of their adoption , technical performance, usability performance, and clinical impact. Two key properties shape the quality of XAI systems; faithfulness, which refers to how accurately an explanation reflects the model's actual reasoning, and plausibility, refers to how understandable and meaningful the explanation is to users. Faithfulness is assessed using objective metrics and plausibility is assessed using subjective metrics. Furthermore, it is important to note that plausibility can be assessed from a multidimensional perspective using explanation evaluation metrics (ensuring AI's reasoning is clear, trustworthy, and helpful) and usability metrics (ensuring the system is intuitive, easy to use, efficient, and practical).This classification helps researchers identify gaps in specific areas of evaluation of XAI methods, ultimately supporting the development of systems that are both interpretable and usable in real-world settings. This review critically synthesises current evidence to address these key aspects to understand the effectiveness of XAI methods in regards to support clinical decision making, improve diagnostic care and enable personalised care. The findings of this RQ2 are structured along three thematic axes: (i) enhancement of decision support, (ii) improvements in diagnostic accuracy, and (iii) enablement of personalised care, with XAI effectiveness influenced by explanation plausibility and faithfulness.

The overall high methodological quality of the included studies enhances the robustness of this review's findings. Mixed-methods and qualitative studies are well-executed ensuring the results are based on data both rigorously collected and appropriately interpreted. While a small subset of studies exhibited methodological weaknesses such as insufficient integration of data types or lack of clarity in reporting ([52], [53], [55], [59], [61]) these were critically appraised and contributed minimally to the thematic synthesis. Their inclusion ensured completeness but did not materially influence the main results. Hence, the synthesis is based on high-quality evidence, strengthening the validity and trustworthiness of the review's conclusions regarding XAI usability and impact in healthcare settings.

3.2.1. XAI Methods by Healthcare Domains and Stakeholder Groups

The synthesis of findings demonstrate the wide adoption of XAI methods across various healthcare domains, including endocrinology (e.g., diabetes care [50]), intensive care medicine (sepsis [51]), pediatrics (neonatology [52]), neurology (ALS, epilepsy [53], [65]), oncology (lung cancer [54]), cardiology (ECG analysis [60], [67]), and others such as nephrology [59], infectious diseases [61], and obstetrics [62]. Healthcare providers are the main stakeholders and beneficiaries like pediatricians, nephrologists etc, aiming to enhance clinical decision-making, diagnostic accuracy, and personalised care. Key use cases range from the development of specialised XAI platforms and dashboards for risk prediction and monitoring (e.g., XAI4Diabetes platform [50], interactive diabetes risk dashboards [70]) to improving trust and usability of clinical decision support systems (CDSS) [57], [63], [68]. The majority employ widely recognised XAI techniques such as SHAP and LIME for global and local interpretability, alongside other methods including Grad-CAM for visual explanations [52], example-based explanations [54], [62], counterfactual reasoning [56], [69], and ontology-based approaches [68]. These diverse methodologies highlight efforts to create explanations that are both meaningful and faithful to the underlying models, supporting HCPs in understanding AI predictions and integrating these tools into clinical workflows effectively (see Appendix B , Table B.1 for more details). Overall, the evidence underscores the growing integration of XAI in healthcare to support clinical decision making, improve diagnostic accuracy and enable personalised care through the following use cases.

a) Enhancement of Clinical Decision Support

XAI methods have been widely adopted across different healthcare domains to enhance CDSS and assist healthcare providers in making more informed, transparent, and trustworthy decisions. Domains such as endocrinology (diabetes care), intensive care medicine (sepsis), neurology (ALS, cognitive disorders), pediatrics (neonatology), and general medicine have integrated XAI techniques like SHAP, LIME, Grad-CAM, and counterfactual explanations to provide interpretable and actionable insights for clinicians ([50], [51], [53], [56], [57]). These explanations not only increase trustworthiness by making AI predictions transparent but also improve the alignment of AI recommendations with clinical expertise, as seen in diabetes risk prediction platforms and sepsis detection tools ([50], [51]). In CDSS contexts, XAI supports medical practitioners by clarifying prescription screenings or diagnosis reasoning, thereby facilitating informed decision-making and enhancing system usability ([57], [63], [68]). While [53] contributes valuable insights, it exhibited methodological limitations such as unclear sampling strategies and potential nonresponse bias that may affect the generalizability of its findings; therefore, its conclusions should be interpreted with caution. Nonetheless, the integration of user-centered interfaces and interactive explanations continues to empower healthcare providers to actively engage with AI tools, promoting greater acceptance and effective utilisation within clinical workflows ([58], [68]).

b) Improvements in Diagnostic Accuracy

XAI techniques contribute significantly to diagnostic accuracy by enhancing AI model reasoning and highlighting relevant clinical features that impact predictions. Studies in neonatology, neurology, cardiology, ophthalmology, and infectious diseases demonstrate that explanations such as Grad-CAM visualizations, SHAP values, and example-based reasoning help clinicians verify and validate AI outputs, improving diagnostic confidence ([52], [56], [65], [66], [61]). For instance, in seizure detection, the use of visual explanation modules reduced false alarms and increased sensitivity, providing clinicians with clearer insights into EEG patterns ([65]). Similarly, ophthalmology models achieved high precision and sensitivity with tailored explanations that assist ophthalmologists in understanding patient-specific risk factors for diseases like age-related macular degeneration ([66]). Moreover, in sepsis prediction, the alignment of AI explanations with known clinical concepts (e.g., leukocyte counts) has improved the accuracy of diagnoses, showing how XAI can bridge model outputs with established medical knowledge ([51]). While not all studies report direct diagnostic accuracy improvements (e.g., oncology lung cancer relapse prediction [54], pediatric neurology [71]), many illustrate better human-AI collaboration and error identification, which indirectly contribute to diagnostic robustness.

c) Enablement of Personalised Care

Personalisation through XAI is important where explanations are tailored to individual patient data and clinical contexts, enabling more customised treatment plans. Local explanation methods (LIME, SHAP), counterfactual reasoning, and interactive dashboards help clinicians understand the unique factors driving AI predictions for each patient, supporting personalised risk assessments and intervention strategies ([50], [59], [62], [70]). For example, diabetes risk prediction platforms utilise data-centric and counterfactual explanations to provide patients and clinicians with actionable, personalised recommendations, improving engagement and health outcomes ([70]). In nephrology and obstetrics, explanations offer insights into individual risk factors for diseases like polycystic kidney disease and gestational diabetes, supporting tailored monitoring and management plans ([59], [62]). Furthermore, some studies highlight the importance of adaptable XAI tools that account for the experience level of the user, ensuring that personalisation of explanations aligns with the needs of both novice and expert clinicians ([71]). Such adaptive explainability promotes better decision-making and patient-centered care, emphasising the role of XAI in bridging AI predictions with personalised clinical workflows.

3.2.2. Evaluation Metrics and Methodologies for XAI Models in Healthcare

The assessment of evaluation of XAI methods in healthcare is multidimensional typically addressing three complementary aspects: technical performance, usability, and clinical impact (See Appendix B Table B.3 for more details).

a) Technical Performance Evaluation

Technical performance measures how well the explanations reflect the model's internal reasoning (faithfulness) with metrics such as accuracy, area under the curve (AUC), sensitivity, specificity, F-score, and explanation fidelity. Quantitative and qualitative metrics are used to ensure reliability, predictive accuracy, and faithful representation of model behavior. Commonly used metrics include classification performance indicators such as AUC (Area Under the Curve), accuracy, sensitivity, specificity, precision, and F-score, often applied with cross-validation and hyperparameter tuning to ensure robustness and reproducibility across multiple datasets and experiments [59, 65, 66]. Explanation-specific metrics such as fidelity (comparing the surrogate to the original model), feature importance agreement, faithfulness, causality, transferability, confidence, and trustworthiness assess how well the explanation reflects the underlying model [50, 51, 52, 53, 59]. Attribution metrics including monotonicity (Spearman correlation), implementation invariance (Jaccard similarity), and weight of advice quantify the alignment between explanations and true model behavior [50, 51]. Computational performance metrics like explanation complexity, hit rate, and computational speed (e.g., ClusteredSHAP's efficiency improvements) also contribute to evaluating feasibility in clinical settings [67, 68]. However, studies highlight that technical evaluations must consider model limitations such as confusion with specific data types, false positives, or under-optimized predictive accuracy, which can affect explanation validity [52, 53, 65]. Additionally, regulatory compliance and unbiasedness are increasingly recognised as critical criteria for trustworthy AI deployment [61]. Statistical methods such as Spearman correlations and p-values help validate the relationships between explanation quality and model accuracy [60]. Importantly, technical metrics are often contextualised with expert clinical feedback or medical literature to ensure clinical relevance and applicability [50, 51].

b) Usability and Interpretability Evaluation

Usability and interpretability are often evaluated through clinician-centered methods, including qualitative feedback, surveys, Likert-scale questionnaires, and think-aloud protocols. These methods increase the understandability, plausibility, and trustworthiness of explanations from the end-users' perspective. It is mostly assessed using mixed methods combining qualitative and quantitative approaches to capture clinician perceptions, cognitive responses, and interaction experiences. Common methodologies include structured surveys, Likert-scale ratings on trust, confidence, comprehensibility, usefulness, and satisfaction; interviews; think-aloud protocols; cognitive walkthroughs; focus groups; and standardised usability instruments like the System Usability Scale (SUS) and User Experience Questionnaire (UEQ) [50, 52, 54, 55, 56, 57, 58, 61,

63, 67]. These evaluations focus on metrics such as interface clarity, interaction efficiency, workflow integration, interpretability, and alignment with clinical mental models [59, 60, 63, 64, 70]. Studies underscore the importance of visualization preferences, textual summaries, and hybrid explanation formats (e.g., combining local/global feature importance and counterfactuals) in enhancing comprehension and user satisfaction [55, 56, 64, 70]. However, persistent challenges remain, including cognitive overload, confusing rationale, information complexity, and sociocultural barriers like fear of AI replacing human judgment [50, 52, 63, 68]. User preferences vary by clinical role and experience level, with novices generally valuing explanations more to build trust, while experts may be more critical or less influenced [60, 62]. Usability improvements are achieved through adaptable, interactive dashboards and contextually linked explanations that better support clinical reasoning and decision-making [64, 70]. Overall, iterative, human-centered design and participatory approaches are essential to tailor explanations effectively to diverse HCPs needs and workflows [57, 58].

c) Clinical Impact Evaluation

Clinical impact assessments,examine how XAI influences diagnostic accuracy, decision-making performance, workflow efficiency, and ultimately patient outcomes. Its evaluation centers on how XAI explanations influence clinicians' trust, decision-making, and integration into clinical workflows, ultimately aiming to improve patient outcomes and safety. Key metrics include behavioral intention to use AI systems, weight of advice (WOA) measuring influence on decision changes, explicit and implicit trust assessments, and task performance indicators such as accuracy, efficiency, and error rates [59, 62, 63, 65, 68]. Evidence shows that explanations can enhance transparency and understanding of AI predictions, facilitating more informed, personalised, and collaborative clinical decisions in areas such as ADPKD risk stratification, gestational diabetes, seizure detection, and neurological diagnosis [50, 51, 59, 62, 65]. However, clinical impact is often limited by issues like over-reliance on AI outputs (including incorrect predictions), misinterpretation of explanations, lack of uncertainty quantification, and incomplete integration into existing workflows [62, 63, 68]. Clinician feedback emphasises the need for explanations that align with established clinical heuristics, provide actionable insights, and foster trust through clear, relevant, and contextualized information [55, 57, 66, 70]. The sociocultural context, user expertise, and prior experience with AI systems critically affect acceptance and reliance on XAI tools, highlighting the importance of personalised explanation strategies and ongoing user education to calibrate trust effectively [60, 71]. Moreover, regulatory compliance, workflow compatibility, and continuous user involvement through iterative design are crucial for ensuring practical utility, safety, and sustainable adoption of XAI in real-world clinical settings [51, 61, 68]. A few studies ([52], [53], [55], [59], [61]) demonstrated slightly lower quality with methodological limitations. As highlighted in the quality assessment, their findings should be interpreted with caution, since they contribute less strongly to the overall conclusions of clinical impact.

Effectiveness of XAI system:

The effectiveness of XAI systems in healthcare is shaped by the interplay between faithfulness and plausibility . It has been highlighted from several studies that grounding explanations in actual model features or domain knowledge enhances faithfulness (e.g., [50], [51], [59], [66], [67]). For instance, the use of knowledge graphs in [50] and gradient-based reasoning in ClusteredSHAP [67] showed high alignment with model logic. Similarly, [65] demonstrated strong faithfulness by mapping SHAP values across time, space, and frequency domains in neonatal seizure prediction. However, high faithfulness does not guarantee user trust or comprehension. In several cases, explanations that were technically accurate failed to be clinically plausible either because of complex language ([50]), irrelevant highlighted regions ([52]), or mismatch with clinicians' expectations ([54], [56], [63]). In contrast, example-based explanations were often rated highly for plausibility due to their intuitive nature ([57], [64], [70]), but they sometimes lacked a faithful representation of the model's internal logic ([57], [71]). This gap can be concerning, as plausible but unfaithful explanations may promote automation bias or reinforce incorrect AI decisions ([60], [62]). Furthermore, the effectiveness of XAI varied depending on clinical context and user expertise. Systems like Local Feature Importance (LFI) in [55] showed higher faithfulness and plausibility in high-risk, data-rich cases, while Global Feature Importance (GFI) was more useful in low-risk, general contexts. However, misapplication of explanation types led to reduced interpretability. This was highlighted in [56], where global visualisations aligned better with clinicians' mental models than patient-specific plots. In some studies, explanations increased influence on decisions without necessarily improving trust or correctness, revealing a disconnect between interpretability and decision support ([63], [68]). Over-reliance on clear but incorrect explanations also raises ethical concerns ([57] , [62]), where users trusted explanations even when the model was wrong. In summary, XAI systems in healthcare show promising potential, particularly when explanations are both technically grounded and user-centered but most systems struggle to attain this . This is due to poor communication design, lack of clinical context, or inconsistent alignment with model reasoning. The findings suggest a pressing need for faithfulness-aware design, clinician-in-the-loop development, and rigorous clinical validation to ensure that explanations truly support safe and effective decision-making.

3.3. Usability and interpretability of XAI

This section presents our findings on understanding interpretability and usability needs of clinicians and patients. More specifically, we seek to answer RQ 2 by gathering evidence on how user studies assess if XAI meets HCPs and patients interpretability and usability needs. The RQ2 is framed as "How do user studies evaluate whether XAI meets the interpretability and usability

needs of both clinicians and patients?" and is broken into two sub questions (a) What study designs, user-centred approaches, and experimental frameworks are used to assess the usability and interpretability of XAI in healthcare (RQ2.1) (b) What challenges are faced by HCPs and patients for real-world adoption of XAI tools (RQ2.2).

3.3.1 Study designs, user-centred approaches, and experimental frameworks

User studies assessing the usability and interpretability of XAI in healthcare employ a diverse range of study designs, user-centered approaches, and experimental frameworks to capture the multifaceted needs of both clinicians and patients (See Appendix C, Table C.1 for more details). These studies aim to evaluate how effectively XAI systems communicate decision-making processes, how understandable and actionable the explanations are, and how well they integrate into clinical workflows or patient decision-making contexts.

Study Designs and Datasets

Most studies use mixed-methods designs, combining quantitative metrics with qualitative feedback to capture both objective performance and subjective user experience ([55], [56], [57], [58], [66], [68], [70]). This approach allows researchers to assess not only system accuracy but also user trust, preferences, and interpretability. Quantitative descriptive designs evaluate measurable indicators such as prediction accuracy, feature importance, or explanation clarity ([53]) whereas qualitative designs often incorporate think-aloud protocols, interviews, focus groups, and thematic analyses to explore cognitive processes and subjective interactions with XAI tools ([54], [64], [67], [69]). The studies reviewed in this work utilise a wide variety of non-imaging datasets reflecting diverse healthcare domains and tasks relevant to XAI applications.Several studies employed well-known clinical datasets, such as the Pima Indians Diabetes Dataset and data from Sylhet Diabetes Hospital for diabetes risk prediction [50], ICU patient data for sepsis prediction [51], and MIMIC-IV for intensive care scenarios [68]. Others focused on specialized cohorts, such as neonatal ventilation data from RWTH Aachen University Hospital [52], ALS progression data from the iDPP CLEF 2022 challenge [53], lung cancer relapse prediction using knowledge graph-enhanced clinical data [54], and brain connectivity data from the ADNI database for MCI classification [56]. Large-scale EHRs were used for cardiovascular risk prediction (e.g., MACE) in [55], while pediatric-specific datasets, such as those from Zhejiang University Children's Hospital [64] and Helsinki University Hospital EEG recordings [65], supported use cases in neonatal and pediatric care. Studies also explored genetic and demographic data for rare conditions like AMD [66] and ADPKD [59]. Several papers [57, 60, 71] relied on survey-based data involving medical professionals or the public to evaluate trust and decision-making in XAI systems. Notably, datasets for some studies were not explicitly mentioned [58, 61, 63, 69], though their clinical contexts were inferable. These datasets collectively span a range of healthcare applications including diagnosis, risk stratification, and

clinical decision support, offering a rich basis for evaluating XAI methods across different data types and clinical settings.

User-Centred Approaches

User-centred methods focus on directly engaging end users clinicians, patients, and developers to understand their interpretability needs and adapt explanations accordingly. Common approaches include Surveys and questionnaires,Think-aloud protocols,Focus groups and semi-structured interviews,Clinician feedback loops and iterative co-design ,Human-AI collaboration studies,User expertise differentiation. Surveys and questionnaires are frequently used to collect structured feedback on trust, usability, and explanation effectiveness ([50], [58], [59], [67], [69]). Think-aloud protocols involve participants verbalising their thoughts while interacting with XAI tools, providing insights into cognitive processes and explanation clarity ([54]). Focus groups and semi-structured interviews gather in-depth insights into user preferences, cognitive demands, and suggestions for improvement ([69], [70], [71]). Clinician feedback loops and iterative co-design methods refine explanations based on domain-specific needs and clinical workflows ([52], [61], [64]). Human-AI collaboration studies investigate how explanations affect trust calibration, decision confidence, and advice-taking ([57], [62], [63]). Finally, user expertise differentiation involves comparing interactions across users with varying clinical experience, helping to tailor explanations to different knowledge levels and needs. ([60], [61]).

Experimental Frameworks and Evaluation Metrics

The rigorous evaluation of interpretability and usability in XAI requires experimental frameworks that integrate explainability techniques with human factors research. Common explainability methods include popular XAI techniques such as LIME, SHAP, Integrated Gradients, and Grad-CAM, which generate interpretable model outputs ([50], [51], [52], [53], [56], [59]). To enhance clinical relevance and clarity, visual explanation tools like Brain Connectivity Plots, Counterfactual Plots, Global Tornado Plots, and Knowledge Graphs are also explored ([54], [56], [67]). Quantitative metrics are applied at multiple levels: model performance metrics such as accuracy, precision, recall, F1-score, and AUROC evaluate the reliability of the underlying AI system ([50], [52], [59], [65]); explanation-specific metrics including faithfulness, fidelity, separability, and concept scores assess the correctness and relevance of the generated explanations ([50], [51], [53]); and human-centered metrics like trust scales, Weight of Advice (WOA), cognitive load, and behavioural intention quantify user trust, reliance, and mental effort during interaction with XAI systems ([55], [57], [62], [63]). Analytical approaches such as ANOVA, correlation, and regression are commonly used to study the effects of different explanation strategies on user experience ([57], [60], [71]), while more rigorous experimental designs, including cross-validation and randomised controlled trials, are

employed to compare various XAI methods and their impact on clinical decision-making ([52], [62]).

Overview on Usability and Interpretability needs :

These user studies use comprehensive methodologies, user-centred approaches, and robust experimental frameworks for assessing different usability and interpretability needs of HCPs and patients. For clinicians, it is essential that AI systems are not only accurate but also user-friendly and seamlessly integrated into existing workflows. A trustworthy XAI system should have these usability features: intuitive user interfaces, seamless integration, and actionable insights. Unlike clinicians, patients do not require technical validations but instead require clear risk assessments, treatment options, and reasoning behind AI-generated recommendations, emphasising the need for simple, understandable, and transparent explanations. This could involve interactive explanations, visual aids, and simplified summaries of AI-generated medical reports to effectively support patient understanding and trustworthiness.

In this review, several themes emerged regarding usability and interpretability needs of XAI tools in clinical settings. Usability findings indicated that most clinicians found AI tools easy to use when they featured simplified language, simplified explanations, clear context, and intuitive interfaces ([50], [64], [68], [69]). For instance, [69] highlighted clinicians' preferences on simplified explanations, with a focus on data-centric visualisations (e.g., trends, raw data) rather than complex feature attributions. It also expressed a need for contextual information and for data to be presented alongside interventions. Techniques like SHAP and LIME, paired with textual summaries, significantly improved usability and understanding of model predictions [59]. However, there were recurring challenges, such as misclassifications, misleading visualisations, and effort required to tailor inputs to clinical contexts ([52], [51], [56]). XAI tools were more usable when they supported role-specific needs (e.g., nurses, residents), provided interactive features, allowed personalisation, and aligned with clinical workflows ([51], [53], [64], [70]). TAP and interface redesigns surfaced hidden usability needs, including desires for "what-if" scenarios, uncertainty quantification, and dynamic cohort-based analysis ([54], [68]). Usability was also enhanced by hybrid explanation approaches and reference-based comparisons, though time and cognitive load remained concerns, especially in high-pressure clinical environments ([55], [67], [65]).

On interpretability, clinicians and patients both preferred clear, context-aware, and data-centric explanations, especially those involving visual aids, counterfactuals, and familiar formats like decision trees or reference-value comparisons ([50], [57], [64], [70], [71]). Explanations that linked data, model features, and outcomes directly improved understanding and trustworthiness, especially when they matched clinical heuristics ([50], [51], [66]). Visual tools such as Global Tornado Plots, reference-value displays, decision trees, and counterfactuals were consistently rated as more interpretable than abstract or feature-only outputs ([56], [57], [71]). XAI tools like

SHAP and LIME improved feature attribution visibility, but some methods like Global Feature Importance lacked patient-level clarity ([59], [67]). Trustworthiness was closely tied to alignment with clinical reasoning, clarity of the model's process, and the ability to challenge or validate AI outputs ([52], [60], [65]). Local Feature Importance (LFI) was often seen as more helpful than global explanations for individual cases but required more cognitive effort ([55], [59]). Clinicians emphasised the need for explanations that supported rather than disrupted decision-making, with experienced users occasionally finding XAI misaligned with their workflow [71]. Explanations improved when they provided causality, aligned with known risk factors, included textual summaries, and allowed interaction or customisation ([53], [59], [64], [67]). Nevertheless, issues related to over-reliance on AI, low-quality or misleading explanations, and challenges interpreting outputs in straightforward tasks hinder effective use ([52], [63], [60], [68]) of XAI systems.

3.3.2 Challenges for real-world adoption and applicability of XAI tools

The adoption of XAI tools in healthcare faces several challenges impacting both HCPs and patients, particularly related to usability, interpretability, trust, workflow integration, and deployment affecting the real-world applicability of XAI tools.

a) **Usability Needs and Preferences of Clinicians vs Patients :** Clinicians require AI systems that are accurate, user-friendly, and integrated into clinical workflows. Trustworthy XAI tools should offer intuitive interfaces, seamless integration, and actionable insights tailored to clinical needs. Simplified language, clear context, and data-centric visualisations (e.g., trends, raw data) are preferred over complex feature attributions ([53], [67], [71],). In contrast, patients need simple, transparent, and understandable explanations, such as clear risk assessments, treatment options, and reasoning behind recommendations. Tools like interactive explanations, visual aids, and summaries help improve patient understanding and trustworthiness ([54], [58]).The differing cognitive and informational needs pose a fundamental challenge in designing XAI tools that effectively serve both groups.

b) **Challenges in Clinician Usability :** Many clinicians find AI tools easier to use when explanations are simplified and contextualised, but challenges remain including misclassifications, misleading visualisations, and the effort required to tailor AI inputs to clinical contexts ([55], [54], [59]). Usability improves when XAI supports role-specific needs (e.g., nurses vs residents), allows personalisation and interaction, and fits into existing workflows ([54], [56], [67],). However, high cognitive load and time pressures in clinical environments limit effective use ([58], [70], [68]). Techniques such as SHAP

and LIME enhance feature visibility and understanding but sometimes lack patient-level clarity or clinical relevance ([62], [70]). Hybrid explanations and reference-based comparisons help but require balancing complexity and cognitive load ([58], [70]).

c) Interpretability and Explanation Preferences

Clinicians and patients prefer explanations that are clear, context-aware, and data-centric, using familiar formats like decision trees, reference-value comparisons, and counterfactuals ([53], [60], [67],). Visual tools such as Global Tornado Plots and Counterfactual Plots are more interpretable than abstract feature-only outputs, helping improve trustworthiness by linking data, model features, and outcomes ([59], [60]). Explanations aligned with clinical heuristics, causality, and known risk factors, combined with textual summaries and interaction/customization options, enhance interpretability and trust ([53], [54], [56], [62], [67], [69], [70]).

d) Challenges in Trustworthiness and Alignment with Clinical Judgement

There is often a misalignment between AI explanations and clinical judgment, leaving clinicians uncertain whether to trust AI outputs or rely on their expertise ([56], [67]). Complex or unclear explanations can cause automation bias, where clinicians may overly rely on AI outputs even when explanations are low quality or misleading ([55], [66]). Data quality issues (missing, biased EHR data) and technical constraints such as input length limits reduce the reliability of AI explanations, requiring manual verification and reducing usability ([54], [55], [67]). Clinicians are skeptical of incorporating subjective measures (e.g., anxiety) in predictive models unless contextualized to their clinical reasoning (). Integration challenges remain due to time constraints, limited AI training, and poor temporal data handling, hindering workflow adoption ([54], [59], [66]). High computational demands and limited visual scalability also reduce real-world usability, especially in resource-limited settings ([54], [56], [58], [67]).

e) Patient-Focused Challenges

AI explanations are primarily designed for clinicians, making them complex and technical, which reduces their accessibility and usefulness for patients ([54], [58]). Lack of patient-oriented explanations and feedback mechanisms impacts patient trust and engagement in shared decision-making ([54], [56]). Patients struggle to interpret AI-driven risk assessments and often lack motivation to change health behaviors due to unclear explanations ([58]). The black-box nature of AI, absence of uncertainty

communication, and skepticism towards subjective patient experiences weaken the provider-patient relationship and reduce trust in AI recommendations ([70], [71]).

f) Deployment Limitations

Across multiple studies, a consistent barrier to the adoption of XAI tools in healthcare is their prototype nature, often tested on small, non-representative samples, limiting generalizability and confidence in findings ([53], [55], [56]). Many XAI systems depend on complex components such as knowledge graphs, which while promising, introduce noise and complexity that can reduce explanation clarity ([54]). Technical challenges such as misclassifications and data preprocessing artifacts (e.g., zero-padding) further undermine reliability and trust ([55]). Moreover, some models lack optimization for clinical performance or fail to undergo rigorous clinical expert evaluation, restricting their readiness for deployment ([56]). These foundational limitations underscore that many XAI tools remain at an experimental stage, highlighting the gap between research prototypes and robust, clinically validated systems.

g) Workflow Integration Issues

Integrating XAI tools into the demanding, time-constrained clinical environment presents significant challenges. Several studies highlight how cognitive load, interface complexity, and language barriers interfere with effective use of AI explanations, sometimes causing frustration or disengagement among clinicians ([57], [61]). Time-intensive processes such as manual input selection and lack of adaptive explanations tailored to specific roles or contexts create additional friction ([54]). Misaligned explanations that do not fit naturally with clinical reasoning can disrupt workflows rather than support decision-making ([55]). Furthermore, limited physician participation in some studies suggests that usability concerns may be under-addressed in real-world settings ([58]). Collectively, these issues emphasize that for XAI tools to be practical, they must seamlessly fit into clinical workflows, reduce cognitive burden, and be tailored to user-specific needs.

h) Post-Deployment Feedback and Applicability Gaps

A notable gap across the reviewed studies is the scarcity of mechanisms for ongoing feedback and iterative improvement post-deployment. Most XAI systems lack real-world validation or longitudinal studies to assess trust calibration over time ([53], [59], [60]). The absence of feedback loops from both clinicians and patients limits opportunities to refine explanations or adapt the tools to evolving clinical contexts ([54]). This deficiency is particularly problematic given the context-specific nature of many studies, which restricts broader applicability ([57]). Additionally, limited patient-centric adaptations undermine patient engagement and shared decision-making, weakening trust in AI-driven

recommendations ([54]). These gaps highlight the critical need for deploying XAI tools within frameworks that support continuous learning, user feedback integration, and customization to ensure sustained usability and impact in real-world healthcare settings.

3.4. Trustworthiness in XAI-generated explanations

This section aims to answer RQ3 to identify the qualitative and quantitative methods used to assess clinicians' and patients' trustworthiness in XAI-generated explanations, a key aspect for the successful integration of XAI into healthcare settings. In this review trustworthiness in AI systems has been categorised into seven key framework dimensions for effective trust evaluation. These dimensions include perceived trustworthiness, cognitive understanding, emotional & behavioural trust, temporal trust dynamics,usability & integration, ethical & regulatory trust and patient-centered trust. Each of them represents a critical aspect of trust evaluation, supported by empirical studies that use both qualitative (subjective) and quantitative (objective) methods to measure trustworthiness in AI (Appendix C,Table C. 1:for detailed analysis).

- **Perceived Trustworthiness:** This dimension focuses on users' overall perception of how trustworthy the AI system is, often reflecting clarity, reliability, and completeness of explanations. Primarily Likert scale surveys paired with qualitative feedback like open-ended questions or interviews are identified as measurement methods. These studies measure user confidence in AI predictions and system transparency, which are essential for trust formulation ([50], [52], [54], [56], [57], [63],[64],[65], [66],[70]).
- **Cognitive Understanding:** This captures how well users comprehend the AI's reasoning and explanations, including the quality and clarity of those explanations ([54], [56], [57], [60],[61],[64],[67],[71]). Think-aloud protocols, usability scores (like Avg.EUS), thematic analysis, Task behaviour, justifications, clinicians' observations and qualitative feedback are some of the identified measurement methods. This is a significant dimension of trust because cognitive understanding is crucial as trust is unlikely if users don't understand how the AI arrives at decisions.
- **Emotional & behavioural Trust:** This dimension examines users' feelings (confidence, reassurance) and behaviours (reliance on the system, intention to use) ([50], [52], [55], [62], [63], [68]). Self-reported confidence, Weight of Advice (WOA), behavioural intention (BI) scales are some of the identified measurement methods. Trust isn't just cognitive; emotional responses and actual behaviours (like using the system) are key to successful clinical adoption.
- **Temporal Trust Dynamics:** This emphasises on how trust evolves over time with repeated interactions or usage [60]. Longitudinal Likert scale ratings to compare initial and final trust is a type of measurement method. This dimension is significant since trust

is dynamic, so it's important to track changes, especially as users gain experience with the system.

- **Usability & Integration:** This evaluates how usable the system is and how well it fits into users' workflows, influencing trustworthiness ([57], [62], [63], [67]). Human-Computer Trustworthiness (HCT) scale, system usability scales, and likelihood-to-use surveys are identified as some of the measurement methods. Even a trustworthy AI may be rejected if it's not usable or doesn't integrate well into existing processes.
- **Ethical & Regulatory Trust:** This includes perceptions related to fairness, reliability, predictability, and adherence to ethical or clinical standards ([63], [71]). Negative Attitude Toward Robots Scale, qualitative feedback about system predictability and fairness are identified as some of the measurement methods. Trust also depends on the system's alignment with ethical expectations and regulatory compliance, especially in healthcare.
- **Patient-Centered Trust:** This dimension reflects trust as it relates to supporting patient care, including how explanations help decision-making ([56], [61], [66], [71]). Open-ended feedback, interviews, and surveys evaluating trust in AI outputs from a patient care perspective. In clinical contexts, trust must extend beyond clinicians to the patient level to ensure acceptance and adherence to AI-guided decisions.

This review highlights several key findings about the assessment of clinicians and patients' trustworthiness.

a) Trustworthiness is self-reported by clinicians based on their improved understanding of AI predictions and confidence in using the system [50,71]. Similarly,[55] assesses trustworthiness through HCPs ratings on system reliability and their feedback on trustworthiness and reliance on AI-generated explanations, comparing global vs. local feature importance.
b) The level of trustworthiness varies between novice and expert clinicians. For eg., [60] and [71] highlighted novice clinicians and general population trustworthiness XAI explanations more than expert medical professionals due to over-reliance on AI.
c) Visual explanations and decision path transparency are often linked with trustworthiness. In [61], trustworthiness is assessed through clinicians' feedback on the transparency of risk factors and the clarity of decision path visualisations. Similarly, [70] emphasises the importance of visual explanations and data transparency as significant factors influencing trustworthiness.
d) Trustworthiness is important for usability and the system's ability to fit into clinical workflows. [61] highlights the need for user-centred evaluation to generate trustworthiness and ensure the usability and usefulness of the system.
e) In some studies, trustworthiness is measured through a combination of cognitive-based evaluations (perceived understandability, reliability) and behavioural indicators

(agreement with AI recommendations, switching decisions, and human-AI performance), as well as through qualitative and quantitative methods to provide a comprehensive evaluation of user trustworthiness in the AI system [57].

f) There are studies reflecting clinicians' overall trustworthiness in explainability over traditional methods [65] by measuring the improvement in decision-making efficiency, interpretability, and confidence levels.
g) This review on patients' perception of XAI-generated explanations is necessary to get an understanding of how clinicians and patients perceive the explanations provided by XAI systems for the enhancement of the design, usability, and performance of AI tools, limiting the assessment of patients' trustworthiness.

4. Discussion

The study has provided valuable insights into the use of explainability in healthcare for non-imaging data, in context of the emerging needs for transparent, trustworthy, and interpretable AI systems that are crucial for effective integration of AI into clinical workflows.This multidisciplinary research has found momentum since 2023, which highlights that the use of XAI is no longer limited to technical domains but actively involves biomedical and clinical experts. This multidisciplinary nature of research on XAI reflects the inherent complexity of healthcare management mechanisms, where diverse specialties ranging from intensive care to oncology and cardiology require tailored explanation strategies to fulfil their specific and targeted needs. The evaluative approaches reveal a significant paradigm shift towards human-centred metrics, beyond traditional objective metrics like accuracy and fidelity. Increasingly, subjective dimensions such as trust, usability, and explanation clarity are being emphasised to better align with the real-world needs and experiences of end-users in clinical settings.This dual focus is essential, as technically accurate explanations are insufficient if they are not understood or trusted by healthcare providers. Furthermore, the distinction between explanation evaluation (e.g., clarity, trustworthiness) and usability evaluation (e.g., intuitiveness, efficiency) underscores the multifaceted nature of XAI effectiveness and the importance of integrating explanations into existing clinical workflows. A key challenge lies in balancing faithfulness and plausibility: technically faithful explanations may be difficult to interpret, while more intuitive ones may misrepresent model reasoning. This gap is further complicated by variability in user expertise and clinical context; what aids a novice may frustrate an expert ([55], [58]). Therefore, to ensure the reliability and usefulness of AI explanations, XAI must combine an accurate representation of model logic with clear, context-sensitive communication. This demands iterative design, rigorous validation, and tailored strategies that reflect the complexities of healthcare decision-making, which invariably makes the study on XAI a multidimensional exercise.

The multi-dimensional nature of perceived trustworthiness in XAI systems reveals that trust is deeply intertwined with both cognitive and effective user experiences. These dimensions, ranging from cognitive understanding and emotional confidence to workflow integration and ethical alignment, highlight that trust is not a monolithic concept but one shaped by diverse, context-sensitive factors. Notably, these dimensions also interact with the persistent trade-off between faithfulness (i.e., the explanation's fidelity to the model's internal logic) and plausibility (i.e., how intuitively reasonable the explanation appears to users). For instance, cognitive understanding hinges on the clarity of explanations, which can suffer if excessive complexity is nurtured for the sake of maintaining faithfulness. Conversely, emotional and behavioural trust may increase with simpler, more intuitive explanations, though these may mask the actual reasoning of the model, reducing faithfulness. Ethical and regulatory trust also depends on faithfulness to transparent and auditable decision logic, while patient-centred trust requires plausible, digestible explanations that support shared decision-making. Thus, trust formation is not only multidimensional but also highly sensitive, as ensuring optimum balance between technical accuracy and user interpretability in explanations is a very delicate and difficult task. This reinforces the importance of tailored, iterative design approaches in XAI studies that explicitly address the contextual and temporal dynamics of trust and carefully navigate the faithfulness plausibility tension to support safe and reliable AI adoption in healthcare. The explanations have shown to increase clinician confidence, particularly when they align with clinical knowledge and present transparent reasoning through visual tools, enhancing trustworthiness on the system. However, the risk of automation bias, especially with certain explanation types like counterfactuals, signals a need for caution; over-reliance on AI without adequate skepticism could potentially jeopardise patient safety. This is reinforced by the observation that expert clinicians tend to be more critical and demand explanations that complement rather than replace their judgment. Usability studies highlight the importance of customisation and interactivity. Different user groups, novices and experts, require varied explanation complexities and formats, from simple visuals to in-depth dynamic insights. Such personalisation is critical for fostering engagement and trust, yet remains underexplored in current XAI tools, particularly for patients. The marginal consideration of patients in these studies represents a significant gap, suggesting the field has yet to fully embrace patient-centred AI design despite patients being ultimate beneficiaries of healthcare decisions. Despite promising advances, several limitations persist. Common methodological issues such as small sample sizes, limited generalisability, and challenges with data quality constrain the reliability and clinical applicability of XAI tools. The predominance of prototype-level tools, often lacking real-world validation and workflow integration, points to a translational gap that impedes sustainable clinical adoption. Moreover, the underexplored areas of user experience, explanation formats beyond textual outputs, and rigorous validation frameworks represent fertile ground for future research. The balancing act between faithfulness (technical accuracy) and plausibility (user comprehension) remains a key challenge. Highly faithful explanations can be overly complex

and reduce usability, while plausible but unfaithful explanations risk misleading users and fostering automation bias. The findings emphasize the necessity of adaptive, user-centred designs that accommodate variability in expertise and clinical contexts.

While prior SLRs on XAI in healthcare have largely concentrated on the technical aspects of explanation methods and model transparency, the novelty of this work lies in its comprehensive focus on user-centred evaluations that address both interpretability and usability needs of healthcare professionals and patients. This study uniquely synthesises mixed-methods approaches, experimental frameworks, and human factors metrics to understand how XAI explanations are experienced and integrated into clinical workflows and patient decision-making. By incorporating both clinician and patient perspectives and emphasising challenges such as trust calibration and workflow integration, our work fills a critical gap in the literature, offering practical insights to advance the real-world adoption of trustworthy and effective XAI tools in healthcare.

5. Conclusion

The research on XAI in healthcare is expanding rapidly, reflecting a broad recognition of the critical need for explainable, trustworthy AI systems to support clinical decision-making. The evidence underscores that the effectiveness of XAI hinges on a delicate balance between technical fidelity and human interpretability, mediated through human-centred evaluation frameworks that incorporate both objective and subjective metrics. Healthcare providers are the primary beneficiaries of XAI innovations, yet the limited focus on patients and other non-clinical stakeholders highlights an important gap that must be addressed to realise truly inclusive AI solutions. Trustworthiness, usability, and clinical relevance are intertwined pillars that dictate the successful translation of XAI tools from research prototypes to everyday clinical practice.

Future research should prioritise robust methodological designs with larger, more representative samples, incorporate diverse explanation formats, and integrate longitudinal, real-world evaluations. Patient-centred design and inclusion must be elevated to ensure AI explanations empower all stakeholders, especially those directly impacted by healthcare decisions. Ultimately, the promise of XAI lies in its ability to enhance clinical decision support, improve diagnostic accuracy, and enable personalised care without compromising the clinician's expertise or patient safety. Achieving this requires ongoing iterative design, rigorous validation, and multi-stakeholder collaboration to develop XAI systems that are both reliable and practically usable, paving the way for their sustainable adoption in healthcare.

Ethical declaration: Ethics approval is not applicable.

Informed consent and clinicians, patient details: The authors declare that the work described does not involve patients or clinicians.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment: This work was supported by the School of Computer Science, University of Sheffield.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Appendices

Appendix A:

Table A.1: Overview of Included Studies by Healthcare Domain, Stakeholders, and Use Cases

Table A.2:XAI Techniques and Their Applications in Clinical Use Cases

Dataset Characteristics and Evaluation Approaches

Overview of Study Assessments Across Performance, Usability, Clinical Relevance, and Key Insights

Evaluation of Faithfulness and Assessment of Plausibility

Appendix C

Table C.1 : User-Centred Approaches and Experimental Frameworks in Healthcare Explainability Studies

Appendix D

Table D. 1: Assessment of Clinician and Patient Trustworthiness in XAI Explanations: Qualitative and Quantitative Methods

Table A.1: Evaluation Metrics

Category	Metric Type	Metrics	Description	StudyId
Objective (Computer-Based)	Model-Based	Fidelity, Separability, Identity, Computational Speed, Explanation Complexity, Hit Rate	Evaluates model reliability, generalization, and fidelity to black-box models	[53], [67], [68]
	Attribution-Based	Monotonicity, Implementation Invariance, Usefulness Score, Decision Impact (WOA)	Measures clarity and importance of features in predictions	[50], [51], [62]
	Example-Based	-	Evaluates how well examples help explain individual predictions	-

Subjective (Human-Centred)	Explanation Evaluation Metrics(Ensuring AI's reasoning is clear, trustworthy, and helpful)	Understandability, Trustworthiness, Satisfaction, Causality, Informativeness, Confidence, Explanation Quality, Perceived Accuracy, Appropriateness, Agreement with Explanation, Actionability, Comprehension, Preference	Measures how clearly and effectively explanations communicate AI reasoning and influence trust and decision-making	[50]–[71]
	Usability Metrics(Ensuring the system is intuitive, easy to use, efficient, and practical)	Ease of Use, Efficiency, Integration into Workflow, Time Required, Cognitive Load, Customizability, Interactivity, Practicality, SUS Score, Effort Expectancy, Behavioural Intention, Social Influence,	Assesses user interaction experience, system fit into clinical workflows, and likelihood of adoption	[50], [53], [57], [61], [63], [68], [70]

		Technical Support Need, IT Proficiency		

Appendix B:

Table B.1: Use cases for each healthcare domain and stakeholder groups for supporting clinical decision making, diagnostic accuracy and enabling personalised care

StudyId	Healthcare Domain	Subdomain(s)	Stakeholder Groups	Primary Beneficiary	Key Use Cases	Common XAI Methods	Clinical Decision Support	Diagnostic Accuracy	Personalised Care
[50]	Endocrinology	Diabetes Care	Healthcare Providers	Medical professionals in diabetes care	Development of XAI4Diabetes platform for transparent diabetes risk predictions	SHAP (Global), LIME (Local), S-LIME	Enhances trustworthiness in AI predictions by providing clear, understandable explanations. Helps clinicians align predictions with their medical expertise.	Provides feature importance explanations and local predictions, aiding in accurate diagnostics. High consistency between feature attributions and true outcomes.	Localised explanations help tailor care plans based on individual patient predictions. Links patient data to external knowledge for more personalised treatment.
[51]	Intensive Care Medicine	Sepsis	Healthcare Providers	Medical professionals in ICU	New metric evaluating AI explanations for CDSS predictions	LR+SHAP, XGB+SHAP, NN+LRP, NN+IG	XAI methods provide feature importance scores to aid clinicians in diagnosis.	Highlighting relevant concepts (e.g., leukocytes) improves accuracy in diagnosing	Identifies relevant concepts and conditions tailored to patient characteristics

							Explanations align with clinical literature.	sepsis. XGB+SHAP improves predictive performance.	(e.g., age, gender).
[52]	Pediatrics	Neonatology	Healthcare Providers, AI Researchers	Neonatal care professionals, AI researchers	Evaluates Grad-CAM for neonatal ventilation classification	Grad-CAM	Visualizes important flow and pressure data points. Confidence value aids trustworthiness of results.	Explains why certain breaths are classified. Helps identify misclassifications and provides insights.	Tailors decision-making based on individual breath data.
[53]	Neurology	Amyotrophic Lateral Sclerosis (ALS)	Healthcare Providers, AI Researchers	ALS clinicians, AI researchers	Comparison of SHAP, LIME, AraucanaXAI for ALS mortality prediction	SHAP, LIME, AraucanaXAI	Helps clinicians understand model predictions. Identifies important features driving predictions.	Explains the model's behaviour for wrong predictions. Identifies which features contribute most to predictions.	Provides insights based on individual patient data. Highlights critical features for specific patients.
[54]	Oncology	Lung Cancer	Healthcare Providers, AI Researchers	Oncologists, AI policymakers	Oncologists assess lung cancer relapse prediction system	Example-based explanations	Assists in adjusting therapeutic guidelines Informs follow-up schedules	No clear improvement in diagnostic accuracy	Not explicitly enabling personalised care

[55]	General Medicine	Electronic Health Records	Healthcare Providers	Clinical decision-makers	Comparison of explanation methods on EHR outcomes prediction	Global Feature Importance, Local (LIME)	Global: Provides general feature importance across patients. Local: Offers personalised insights for individual predictions.	No direct improvement in diagnostic accuracy mentioned.	Local Feature Importance performs best for patients with high amounts of data (Group G2). This is because such patients have more observations that significantly affect the prediction, which can't be easily captured by other methods. Therefore, Local Feature Importance provides more personalised explanations for these patients.
[56]	Neurology	Cognitive Disorders (MCI, AD)	Healthcare Providers	Neurologists	Interfaces integrating AI for diagnosing MCI and Alzheimer's	Global Tornado Plot, Brain Connectivity, Counterfactual, SHAP	Provides interpretable explanations of AI decisions. Enhances clinician confidence in AI outputs.	Uses brain connectivity to classify MCI and HC. Identifies patterns not apparent in traditional methods.	Tailors diagnostics based on individual brain connectivity.

[57]	Medicine	CDSS	Healthcare Providers	Medical practitioners	Impact of four XAI classes on trustworthiness in CDSS	Local, Example-based, Counterfactual, Global explanations	XAI methods were tested in a CDSSsto assist medical practitioners. Participants performed Human-AI tasks where they used AI recommendations for prescription screening, helping them make informed clinical decisions.	XAI explanations enhanced Human-AI task performance and accuracy. Example-based and Counterfactual explanations were perceived as more understandable and helped improve diagnostic accuracy.	Users could customize explanations and engage with interactive tools, potentially leading to more tailored care. Participants could steer the explanation process to suit their task requirements, aiding in personalising the decision-making process.
[58]	Medicine	Pediatrics	Healthcare Providers	Pediatricians	Designing CDSS explanations for pediatric diagnosis	Various UI designs (e.g., supporting/contradicting factors)	Enhanced understanding of system reasoning. Detailed comparisons of diagnoses. Trustworthiness calibration through system performance insights. Case-based information to evaluate potential diagnoses.	Confusion matrix for system performance (UI 12). Counterfactual statements (UI 8). System performance indicators to assess accuracy. Clinicians rated UI designs highly for informing accuracy.	Personalised feedback is necessary for clinicians, as interface designs can vary in importance depending on the context (e.g., disagreement among colleagues or parents). Training and system familiarity are

									essential for ensuring that clinicians receive the most useful information.
[59]	Nephrology	Polycystic Kidney Disease	Healthcare Providers	Nephrologists	Predicts kidney enlargement in ADPKD with XAI	LIME, SHAP, ChatGPT for text explanations	Enhances model transparency for doctors Helps understand prediction rationale Aids in informed decision-making	Provides insights into feature contributions Validates model predictions, reducing reliance on black-box models	Identifies key features affecting predictions Helps tailor patient treatment based on personalised insights
[60]	Cardiology	Electrocardiography	Healthcare Providers	ECG readers	XAI in AI-supported ECG reading	Textual explanations	Provides diagnostic support with AI suggestions. Textual explanations motivate the AI's diagnostic advice.	AI accuracy: 70% compared to ECG Wave–Maven gold standard. Assists in complex cases, especially for novices.	No direct mention of personalised care.
[61]	Infectious Diseases	Hospital-onset Bacteremia	Healthcare Providers	Infectious disease clinicians	User-centered XAI for HOB risk prediction	Literature-based XAI methods	Generates trustworthiness in the system Helps clinicians understand decision paths Enables tailored	Ensures unbiased results Aims to identify treatable risk factors correctly	Clinicians can plan tailored interventions based on factors driving decisions

							interventions based on identified risk factors		
[62]	Obstetrics	Gestational Diabetes Mellitus	Healthcare Providers	Obstetricians, dietitians	Comparison of feature contribution and example explanations in GDM prediction	SHAP, Explanation by example	Provides transparent explanations of model predictions to aid healthcare practitioners in making informed decisions.	By offering clear, understandable explanations of GDM risk prediction, it helps practitioners verify or adjust their risk estimates.	The use of feature contribution and example-based explanations provides personalised context to the practitioner, allowing them to compare the patient's case to similar ones and understand how individual features contribute to the predicted GDM risk.
[63]	Medicine	CDSS	Healthcare Providers	Healthcare providers	Impact of explanations vs no explanations in CDSS	Dr.XAI (with explanation), Dr.AI (without)	Dr.AI: Provides only a suggestion, but healthcare providers remain solely responsible for the final decision. Dr.XAI: Provides an explanation of the	Dr.XAI: By offering an explanation, it potentially aids in refining the decision-making process by revealing the conditions that most influenced the algorithm's	No direct mention of personalised care.

							algorithmic suggestion, potentially helping healthcare providers understand the reasoning behind the recommendation.	decision.	
[64]	Healthcare Analytics	Predictive Health Technologies	Healthcare Providers	Clinical teams	VBridge tool integrating ML explanations in clinical workflow	SHAP, feature hierarchy, visual context	Explains predictions with context. Improves team communication. Identifies high-risk factors early. Helps junior doctors in making more accurate diagnosis	Reference values aid in interpretation. Reduces decision-making mistakes. Highlights high-risk patients.	Reference values tailored to cohort. Identifies patient-specific risk factors.
[65]	Neurology	Epilepsy, Seizure Detection	Healthcare Providers	Neurologists	XAI4EEG for seizure detection with visual explanations	SHAP, SHAP1D, SHAP3D	Helps Validate Predictions: Supports clinicians in validating seizure detection predictions by providing visual explanations. Increases Confidence	Higher Sensitivity: 1D-CNN achieved 86% sensitivity, indicating good detection of seizures. False Alarm Reduction: The use of the explanation	Customization of Explanations: The module accounts for individual EEG feature contributions, aiding clinicians in personalising care based on each neonate's

							and Trustworthiness: The explanation module improved participants' confidence and trustworthiness in the model predictions.	module allows clinicians to spot errors or misclassifications (e.g., false positives) more effectively.	unique EEG pattern.
[66]	Ophthalmology	Age-Related Macular Degeneration	Healthcare Providers	Ophthalmologists	XAHN for AMD diagnosis using non-visual data	XAHN Explainer, SPE-AHN algorithm	Provides global and local explanations of AMD predictions. Survey with clinicians to assess the usefulness of explanations. Developed an interactive explainer tool based on expert feedback.	AHN Model Accuracy: 98.81% (Avg: 98.45±0.23%). Sensitivity & Specificity: 98.91%. Precision: 98.72%. F-score: 98.80%.	Allows clinicians to test individual cases and get tailored explanations. Provides case-specific insights on risk factors influencing AMD diagnosis.
[67]	Cardiology	Electrocardiography	Healthcare Providers, Patients	Cardiologists, ECG patients	Enhancing explanation of 12-lead ECG classification	GradientExplainer, ClusteredSHAP, DeepExplainer	Provides faster and clearer explanations for real-time medical decisions. Helps clinicians interpret	ClusteredSHAP maintains or improves explanation quality while reducing computational time.	No direct mention of personalised care.

							AI-generated ECG results effectively.		
[68]	CDSS + Healthcare Analytics	Explainable AI in decision-making	Healthcare Providers	Healthcare providers	XAI system to improve trust and interaction with CDSS	Doctor XAI, Ontology-based, Rule extraction, Model mimicry	Interpretability of AI Predictions Human-Centred Interface (interpretable explanations with patient's clinical history) Trustworthiness in AI (increased trustworthiness via explanations) Behavioural Intention (encourages AI use through explanations)	Synthetic Neighborhoods (improves explanation fidelity) Ontology-Based Perturbation (enhances decision-making fidelity)	Sequential Data Handling: It processes patient histories over time, considering sequences of events to provide context-specific explanations. Multi-Label Predictions: It predicts multiple conditions simultaneously, offering comprehensive care by addressing various health aspects for a patient. Contextualized Explanations: The system generates tailored explanations based on

									individual patient data, ensuring the AI's reasoning aligns with the patient's unique conditions.
[69]	Healthcare Analytics	Predictive Health Technologies	Healthcare Providers	Health experts	Predictive health tech to aid experts understanding	Counterfactual, Feature Importance	Transparent health dashboard for experts CF gives concrete intervention suggestions FI highlights key health parameters Data-centric explanations for informed decisions	Identifies missing relevant parameters (e.g., sleep, workload) Assesses appropriateness of model parameters CF highlights problematic parameters for refinement	The study highlights the importance of incorporating situational data to provide better context for health experts. Participants emphasised the need for explanations that not only highlight negative health parameters but also incorporate positive progress in patient health and recovery. Additionally, integrating patient-specific contextual data, such as work environment and lifestyle

									factors, was suggested to improve the relevance and accuracy of intervention planning.
[70]	CDSS + Healthcare Analytics	Diabetes Risk Prediction	Healthcare Providers, Patients	Diabetes patients, clinicians	Interactive dashboard for diabetes risk monitoring	Data-centric, Feature importance, Counterfactual	Data-centric explanations help HCPs identify high-risk factors. Feature-importance explanations assist in suggesting actions for risk reduction. Counterfactual explanations provide actionable recommendations for patients.	Data-centric explanations offer detailed insights and risk comparison. Feature-importance explanations highlight significant risk factors. Counterfactual explanations give personalised, actionable advice.	Data-Centric Explanations: Personalised health insights, allowing HCPs to compare a patient's data with population trends for more tailored care. Counterfactual Explanations: Provides actionable, patient-specific recommendations based on their unique health data, guiding informed decisions. Simplified Visuals: Makes complex health

									information easy to understand, especially beneficial for older patients or those less familiar with managing their health.
[71]	Neurology	Pediatric Neurology	Healthcare Providers	Neurologists	AI-based DSS in pediatric neurology decision-making	Decision Tree, Counterfactual, Feature Importance, Case-based explanations	XAI helps less-experienced neurologists improve their performance when perceived as more explainable. Expert neurologists may find XAI disruptive if the explanations are perceived as highly explainable, affecting their decision-making.	No clear improvement in diagnostic accuracy with XAI compared to no-XAI conditions. Some improvement for less-experienced participants, but expert neurologists' performance may degrade with more explainable XAI methods.	XAI tools need to be adaptable to users' experience levels for better support. Personalisation is essential to align explanations with the user's needs and experience.

Table B.2: Evaluation metrics and methodologies used to assess the technical performance, usability, and clinical impact

StudyId	Dataset	Dataset Description	Evaluation Metrics	Evaluation Methodology	Evaluation Methods Used	Technical Performance	Usability / Interpretability	Clinical Impact	Key Insights
[50]	Pima Indians Diabetes Dataset; Early-stage diabetes risk dataset from Sylhet Diabetes Hospital	768 instances (Pima), 520 individuals (Sylhet); female patients, diabetes risk prediction	- Attribution Metrics: Monotonicity (Spearman correlation), Implementation Invariance (Jaccard similarity). - Usability Metrics: Trust, User-friendliness, Confidence, Alignment with Medical Judgment, Need for Support (5-point Likert scale).	Mixed Methods	Survey, participant feedback, Spearman correlation, Jaccard similarity, attribution comparisons	XAI4Diabetes app links ML models, datasets, and ontologies to provide global/local diabetes risk explanations.	Users find explanations helpful but want simpler language, more context, and better model details; interface improvements planned.	Increases trust by making AI outputs transparent, aiding clinical decision-making in diabetes risk.	User-centric, multifaceted explanations improve interpretability; needs more user studies, feedback loops, and patient-focused enhancements for wider adoption.
[51]	ICU patient data for sepsis prediction and Early Warning System (EWS)	Clinical patient data; vital signs and lab results; exact dataset unspecified	- Attribution Metric: Usefulness Score (clinical concept alignment), Concept Scores. - Quantitative: Faithfulness score, AUROC comparison. - Qualitative:	Mixed Methods	Qualitative validation with clinical literature, concept & faithfulness scores, AUROC, usefulness scores	Novel data-type-independent metric evaluates XAI explanations using biomedical knowledge graphs (KGs), supporting multi-models and multi-modal	Aligns with clinician expectations on clinical concept granularity; requires expert effort for mapping concepts and identifying foils, affecting ease of use.	Boosts clinician trust by linking explanations to clinical KGs; sepsis case aligns with medical literature, aiding AI adoption.	Explanation usefulness is context-dependent; KGs proxy clinical meaning; XAI design must balance automation and domain input; temporal data handling needs

			Comparison with literature, interpretation of scores.			inputs.			improvement; shifts toward user-centered AI evaluation.
[52]	RWTH Aachen University Hospital neonates ventilation and vital parameters	18 patients; 6304 breath segments; invasive mechanical ventilation data sampled at 125 Hz	- Explanation Evaluation: Trustworthiness, Causality, Transferability, Informativeness, Confidence, Fairness, Appropriateness, Predictability, Consistency. - Usability: Practicability in clinical use. - Likert-scale ratings, questionnaire, group comparison.	Mixed Methods	Questionnaires (Likert scale), verbal feedback, statistical comparison between experts & developers	Poor classification performance (especially spontaneous vs. mechanical breaths) due to CNN and padded data issues.	Visual explanations often confusing, lacking clear rationale; clinicians want explicit, coherent explanations.	Not suitable clinically due to risks of misinterpretation; recognized potential for research use.	Experts critically assess AI; trust erodes if explanations contradict clinical logic; zero-padding misleads visualizations; larger user studies needed to refine explanations.
[53]	iDPP CLEF 2022 ALS progression prediction dataset	1756 training, 494 test patients; static and time-dependent clinical and spirometry variables	- Model-based Metrics: Fidelity (surrogate vs. original model), Identity, Separability. - Efficiency: Explanation generation time. -	Quantitative	Fidelity, identity, separability, time metrics (no clinician involvement)	SHAP outperforms others in explainability but model unoptimized; temporal ALS dynamics under-addressed.	SHAP and ARAU better align with features; ARAU offers controllable complexity; LIME less aligned; usability limited without	Limited clinical relevance due to no expert involvement and unoptimized models; future utility depends on stakeholder inclusion and	Local explanations key in human-AI disagreement; human-in-the-loop essential; end-user and ethics involvement needed in evaluation.

			Quantitative only, no clinical user feedback.				human tuning.	clinical alignment.	
[54]	Clinical data from 1,348 early-stage lung cancer patients	Integrated into a knowledge graph for lung cancer relapse prediction	- Explanation Evaluation: Confusing/Not, Overwhelming/Not, Complete/Missing, Useful/Not, Misleading/Clear. - Usability: Ecological Validity (real-world applicability). - Methods: Think-aloud protocol, thematic analysis, binary responses.	Qualitative	Think-aloud protocol, thematic analysis, binary-response questions	Predictive relapse score useful but lacks uncertainty quantification and dynamic "what-if" features; explanation fidelity limits robustness perception.	Mixed feedback; clinicians find explanations confusing due to overload and poor clarity; TAP method effectively surfaces usability issues.	Limited but promising; clinicians willing to act but lack confidence due to poor transparency and interpretability.	TAP excels at expert feedback elicitation; credibility and utility perceptions vital; clinicians favor heuristics-mirroring explanations; early expectation alignment improves design.
[55]	Electronic health records (EHR) from 80,000+ hospitalizations at KAGes	Used to train Random Forest model predicting Major Adverse Cardiac Events (MACE)	- Explanation Evaluation: User Trust and Reliance (UTR). - Methods: Interviews with clinicians, Likert scale comparison across	Mixed Methods	Interviews, Likert ratings from HCPs	Global Feature Importance (GFI) is efficient but non-specific; Local Feature Importance (LFI) detailed but computationally heavy;	Mixed usability depending on data volume; GFI easier with sparse data, LFI better for complex cases but cognitively demanding;	Conditional: GFI sufficient for low-risk; LFI/hybrid needed for high-risk to improve trust and understanding.	One-size-fits-all XAI ineffective; explanation choice must be context-aware; presentation order biases trust; first healthcare user study

			explanation types.			hybrid method adapts explanation to data/risk.	presentation order affects trust.		comparing GFI and LFI.
[56]	Brain connectivity data from ADNI database	Used to distinguish Healthy Control vs Mild Cognitive Impairment (MCI)	- Explanation Metrics: Comprehension, Trust, Satisfaction. - Usability: IT proficiency, ease of using visualizations. - Methods: Open feedback, Likert ratings, comprehension correctness.	Qualitative	Open-ended feedback, Curiosity Checklist, free-form comments	Diverse visualizations for neurological diagnosis; Global Tornado Plot most effective; Brain Connectivity and Counterfactual Plots need refinement and interactivity.	Clinicians find Tornado Plot intuitive; other plots less clear; high interest but varied comprehension; need improved design and training materials.	AI seen as a diagnostic aid complementing expertise; cautious optimism due to integration and training gaps.	Visualizations must match clinical mental models; domain-specific training crucial; AI complements, does not replace clinicians.
[57]	Survey data from 41 medical practitioners assessing trust calibration in human-AI decision-making	Experts recruited from three organizations; potential selection bias noted	- Explanation Metrics: Understandability, Reliability, Technical Competence (via Human-Computer Trust scale). - Usability: Workflow fit, cognitive load, customization. - Methods: Interviews,	Mixed Methods	Interviews, thematic analysis, trust scale (Madsen & Gregor), ANOVA, Wilcoxon, Friedman tests	Diverse XAI types improve Human-AI team accuracy; explanations fail to help detect AI errors; technical performance varies with task and explanation.	Example-based and Counterfactual explanations more understandable and engaging; poor interpretability leads to skipped explanations; interactive, tailored methods needed.	Insights relevant beyond medicine; explainability affects engagement, not reliability judgment; trust calibration must consider cognitive demands.	Understandability drives trust and engagement; one-size-fits-all ineffective; cognitive biases risk over-reliance; co-design approaches essential.

			repeated-measures ANOVA, Likert, post-hoc tests.						
[58]	Not explicitly mentioned	—	- Explanation Metrics: Clarity, Trustworthiness, Helpfulness. - Usability: Understandability, Efficiency, Practicality. - Methods: Clinician free-text, Likert ratings (median, min/max).	Mixed Methods	Free-text feedback, clinician Likert ratings (usability & trustworthiness)	Supports iterative XAI prototype development with objective contextual evaluation; detects undesired effects early; enables longitudinal trust studies.	Continuous end-user involvement generates user requirements and multi-modal UI design patterns improving understandability, personalisation, and engagement.	Applied in child health CDSS; addresses clinician needs reducing false positives/negatives; supports trust calibration and safer decisions.	Multidisciplinary, user-centered design crucial; explanations help users learn, predict AI, calibrate trust; combining subjective/objective evaluation needed; personalisation and multi-modal explanations essential; design patterns generalizable across domains.
[59]	Polycystic Kidney Disease Outcomes Consortium (PKDOC) dataset	1779 patients with autosomal dominant polycystic kidney disease (ADPKD)	- Explanation Metrics: Trustworthiness, Causality, Transferability, Informativeness, Confidence. - Usability: Visualization	Mixed Methods	Surveys, interviews, open-ended feedback, Likert scales	Robust ML with 63 experiments; strong AUC on imbalanced data for high-risk ADPKD classification.	XAI (LIME, SHAP) with visual & textual summaries improved user understanding and explainability perception.	Provides clearer, interpretable risk profiles beyond traditional metrics; supports informed clinical decisions.	Combining multiple ML models & imbalance handling boosts reliability; human-centered evaluation crucial; focus on building

			preferences, interface clarity. - Methods: Surveys, interviews, Likert scales.						trust needed.
[60]	ECG reader responses (44 readers) evaluating trust and explanation quality in AI-driven ECG diagnostics	1352 responses; human-first and AI-first protocols	- Explanation Metrics: Comprehensibility, Appropriateness, Utility, Explanation Quality, Dominance (influence on decisions). - Usability: Efficiency, interaction protocol effects. - Methods: Interviews, Spearman correlation, p-values, effect size.	Mixed Methods	Open-ended surveys/interviews, Spearman correlation, effect size metrics	AI accuracy ~70%, higher than average readers; human-first protocol increased trust but didn't affect final trust/explanation quality much.	Explanation quality correlated with trust and decision dominance, especially for novices; novices valued explanations more.	Explanations enhance trust and decision changes but risk misleading decisions; findings limited to ECG context.	Trust varies by user expertise; correctness of AI classification more critical than explanations alone; explanations sometimes misleading.
[61]	Clinical data related to risk factors for Hospital-Onset Bacteremia (HOB)	Dataset unspecified	Explanation Evaluation: Correctness, Transparency, Unbiasedness, Regulatory Compliance.Usability: Information Gain, Time	Mixed Methods	SUS (System Usability Scale), interviews, user-centred evaluation	Compliance with data protection and unbiased models emphasized.	Users need to understand risk factors and decision paths to build trust and plan interventions.	Explainability aids clinicians with useful info and time savings; supports tailored clinical decisions.	User-centered evaluation essential for trust and usability; integration into workflow critical; promising but requires

			Saving, System Usability Scale (SUS), Workflow Integration.Methods: Clinician interviews, open-ended surveys, SUS scores, time measurements.						adaptation.
[62]	Pregnancy Exercise and Nutrition Research Study (PEARS) dataset	Predominantly white Irish women	Attribution Metric: Weight of Advice (WOA) – influence of explanation on decision change.Explanation Evaluation: Clarity, Validity, Clinical Applicability, Preference for explanation type.Usability: Inclination to use, Interpretation Ease, Integration Feedback.Methods: Categorical	Mixed Methods	Survey (preference), thematic analysis, WOA formula	CDSS with feature contribution or example-based explanations influenced decisions; no significant difference between methods.	Practitioners preferred feature contribution; preferences varied by role; inclination influenced advice-taking.	Explainable CDSS shows promise for GDM risk prediction; over-reliance on incorrect predictions is a risk.	User role affects explanation preference; training needed to improve trust; over-reliance challenge remains critical.

			preference questions, thematic analysis, WOA formula, Wilcoxon Signed-Rank Test.						
[63]	Clinical data predicting acute myocardial infarction (MI) using AI-based decision support system (DSS)	—	Explanation Evaluation: Perceived Quality, Trust, Satisfaction, Relevance.Usability: Confidence Shift, Behavioral Intention, Ease of Interaction.Methods: 5-point Likert scales, Think-aloud, Spearman correlation, Wilcoxon tests.	Mixed Methods	Interviews, think-aloud, cognitive walkthrough, Spearman, Wilcoxon, Behavioural Intention (BI)	AI suggestions accurate for acute MI; explanations increased advice influence; error analysis pending.	Explanations increased influence but often unsatisfactory and confusing, especially for novices.	Helped novice mistake prevention and collaborative decisions; sociocultural barriers (fear of replacement) noted.	Explanations raise implicit trust but not explicit confidence; need user-focused design; future work on errors and context needed.
[64]	Pediatric Intensive Care (PIC) Database, Zhejiang University Children's Hospital	Pediatric cardiac surgery patients; demographics, surgery, vital signs, labs, diagnoses	Explanation Evaluation: Clarity, Trustworthiness, Helpfulness (behavioral indicators).Usability:	Mixed Methods	Clinician rating (clarity, trust), case study observation	Links ML explanations with patient records; supports forward/backward analysis; some scalability/dat	Reference-value & hierarchical features enhance interpretability; visual context aids understanding	Reduces blind spots; supports diagnoses & communication; boosts confidence; potential bias mitigation.	Contextual explanations preferred; data quality and biases challenge generalizability; standards needed

			Interface Intuitiveness, Interaction Efficiency, Workflow Practicality.Methods: Case study observations, clinician subjective feedback.			a quality issues.	.		(FHIR).
[65]	Neonatal EEG recordings from Helsinki University Hospital	79 neonates; 1379 seizures; 2640 intervals (350 seizure, 2290 normal)	Explanation Evaluation: Trust, Confidence, Interpretability (self-reported) .Usability: Validation Time.Methods: 5-point Likert scales, timestamp measurements, self-assessment.	Quantitative	Self-reported Likert ratings, timestamps for task duration	1D-CNN showed high sensitivity (86%) and specificity (97.55%) vs. 3D-CNN; hybrid approach balanced strengths.	Explanation module reduced validation time and boosted confidence/trust; hybrid explanation enhanced confidence but not speed.	High sensitivity and low false alarms critical for seizure treatment; explanation module supports decisions and trust.	Domain knowledge integration enhances explainability; hybrid models reveal expert/model disagreements; needs clinical validation.

[66]	Case-control dataset of 256 Mexicans with Age-related Macular Degeneration (AMD) and controls	Includes demographic, clinical, and genetic SNP data	Explanation Metrics: Understanding, Satisfaction, Sufficiency, Completeness, Actionability, Accuracy & Reliability, Trustworthiness. Usability Metrics: Efficiency, Integration into Workflow, Practicality, Intuition. Qualitative: Open-ended clinician feedback. Quantitative: Verbal rating scale (Completely, Mostly, Needs Improvement) on explanation clarity, usefulness, actionability.	Mixed Methods	Verbal rating scale (clarity, usefulness), clinician feedback	AHN model achieved ~98.5% accuracy with strong metrics; robust cross-validation.	XAHN explainer provides global/local explanations; mostly understood but some usability improvements needed.	Highlights known risk factors; supports personalised AMD diagnosis; clinicians value tailored tools.	Combining demographics /genetics improves explainability; human-centered evaluation vital; tailored explainers outperform generic.

[67]	China Physiological Signal Challenge 2018 (CPSC2018) ECG dataset	8 abnormal ECG types and 1 normal ECG type	Model Metrics: Average computational speed, computational complexity. Explanation Metrics: Understandability, Informativeness, Clearness, Usefulness, Accuracy/Relevance, Trustworthiness. Usability Metrics: Avg. Explanation Usability Score (1–7 scale). Qualitative: User experience feedback, interviews. Quantitative: Avg. Explanation Usability Score, Time per sample, Likert scale ratings.	Mixed Methods	UEQ, Avg. Explanation Usability Score (EUS), interviews, Likert, time per task	ClusteredSHAP reduces computation by ~55%, maintains or improves explanation quality; faster than alternatives.	Higher Explanation Usability Scores; less noisy visuals improve clarity.	Faster, reliable explanations aid real-time cardiology decisions and outcomes.	Feature selection via clustering optimizes SHAP; critical for clinical deployment; future work to extend method.

[68]	MIMIC-IV database	ICU patient electronic health records	Model Metrics: Fidelity to Black Box, Hit rate, Explanation Complexity. Explanation Metrics: Explanation Satisfaction, Perceived Explanation Quality, Implicit Trust (Weight of Advice), Explicit Trust. Usability Metrics: Behavioral Intention, Performance & Effort Expectancy, Attitude, Social Influence, Facilitating Conditions, Confidence in Estimate. Qualitative: Interviews, open-ended surveys, observations. Quantitative: Likert scales, task metrics,	Mixed Methods	UTAUT/TAM questionnaires, interviews, task metrics (time, errors), WOA	Explanations increased AI influence (WOA).	Users liked transparency but struggled with complexity; usability improved with simpler, progressive disclosure.	No improvement in confidence or self-reported trust; better explanations correlated with willingness to use.	Explanation quality crucial; poor explanations hurt trust; implicit trust can rise without explicit awareness—automation bias risk.

			Weight of Advice, confidence scales.						
[69]	Not explicitly mentioned	—	Explanation Metrics: Understandability, Practicality, Explanation Preference (CF vs FI). Usability Metrics: Ease of Use, Satisfaction with design. Qualitative: Focus groups, thematic analysis, open-ended feedback. Quantitative: Preference count (CF vs FI).	Mixed Methods	Focus groups, thematic analysis, method preference	Explanations helped identify model gaps but limited for intervention planning.	Users preferred simple, data-centric explanations with trend visuals; counterfactuals viewed as complex/negative.	Aided intervention planning; stressed need for contextual and raw data.	Positive trends should be highlighted; subjective data best as context; expert input valuable for model refinement.

[70]	Electronic health records from five Slovenian primary care institutions	Blood glucose, BMI, waist circumference, age, gender, FINDRISC diabetes risk score	Explanation Metrics: Understandability, Usefulness, Trustworthiness, Actionability, Clarity of Reasoning. Usability Metrics: Ease of Use, Efficiency, Interactivity, Workflow Integrability. Qualitative: Thematic analysis, interviews, open-ended responses, task observations. Quantitative: Task completion rates, Likert-scale responses, justification correctness, time metrics, interaction frequencies.	Mixed Methods	Interviews, thematic analysis, behavioural observation, task completion metrics	Interactive dashboard combining data-centric, feature-importance, counterfactual explanations; supports real-time what-if (offline predictions).	High usability; color-coded visuals and interactivity favored by HCPs and patients.	Improves patient monitoring, communication; aids risk factor identification and motivates behavior change.	Combining explanation types enhances trust and utility; simpler designs needed for older patients.

[71]	Survey data from medical professionals and general population recruited via mailing lists and Amazon Mechanical Turk	Ages 18-65, English speakers, USA; child neurology focus	Explanation Metrics: xAI Performance Improvement, Perceived Explainability, Understanding (−100 to 100), Agreement (−100 to 100). Usability Metrics: Objective/Inappropriate Compliance, Inappropriate Reliance, Trust (Negative Attitude Toward Robots), Anthropomorphism & Social Intelligence (Godspeed scales). Qualitative: Interviews and surveys. Quantitative: Performance scores, scale-based self-reporting, compliance metrics, regression/stat	Mixed Methods	Interviews, surveys, trust/compliance/self-reported understanding; regression analysis, ANOVA	xAI improved performance for less-experienced neurologists but degraded for experts; no overall gain over no-xAI DSS.	Explainability perception varied by method and user expertise; decision trees rated more explainable.	xAI affects compliance and reliance differently by experience; experts showed decreased trust and social competence ratings.	Benefits and harms depend on user expertise; personalisation and alignment with decision-making essential; handcrafted explanations control confounds but limit generalisability.

			istical analysis (ANOVA, logistic regression, etc.).					

Table B.3: Evaluation of Faithfulness and Assessment of Plausibility

Study ID	Evaluation of Explanation Faithfulness	Assessment of Explanation Plausibility	Usability Insights
[50]	Explanations grounded in model inputs and training data; technically faithful.	Helpful and understandable but needs simpler language and more context for better credibility.	Trust in AI predictions increased; easy navigation; user confidence measured; alignment with medical opinion important; some need for technical support reported.
[51]	Faithful via model-derived feature importance validated against clinical knowledge.	Plausible through alignment with clinical concepts, but noisy knowledge graphs reduce reliability.	Usability metrics like trust, ease of interpretation, and user satisfaction not explicitly measured; qualitative feedback absent.

[52]	Low to mixed; some highlights align, others do not reflect true reasoning (e.g., padding).	Inconsistent; some illogical emphasis, reducing clinician trust.	Focus on overall practicality and ease of use in clinical settings.
[53]	Moderate; SHAP/ARAU explanations reflect model in correct cases, less so in incorrect predictions.	Inconsistent; plausible in successful cases but divergent in others, reducing credibility.	Time required to generate explanations tracked as a usability factor.
[54]	Weak; example-based explanations fail to clearly represent model logic.	Variable; example presence aids plausibility but overload/confusion lowers it.	Ecological validity assessed by likelihood of use and recommendation in real work settings.
[55]	Stronger in local feature importance (LFI) capturing patient-specific drivers; weaker in global feature importance (GFI) for individuals.	Higher plausibility in context-aware uses; LFI more plausible for high-risk patients; misuse lowers it.	No usability data reported.
[56]	Strong faithfulness in Global Tornado Plot; patient-specific plots less transparent and less faithful.	Plausibility higher when visualizations align with clinician knowledge; skepticism if logic unclear.	Participants rated visualization understandability and IT proficiency; usability tied to visualization use.

[57]	Not explicitly measured; understandable explanations may not reflect true model logic (over-reliance risk).	High for simple, familiar explanation types, but plausibility doesn't ensure correctness.	Concerns about time constraints and cognitive load; explanations may be overwhelming; customization of explanation features requested; workflow integration challenging.
[58]	Not explicitly addressed; focus on user understanding and trust implies some faithfulness.	Designed for understandable and believable explanations supporting calibrated trust.	Usability assessed through understandability, efficiency, and practicality in clinical workflows.
[59]	Faithful by using LIME, SHAP to represent true feature importance.	Textual/visual explanations increase plausibility and acceptance by aligning with clinical knowledge.	Implied focus on intuitiveness and efficiency; participant feedback used to improve visualizations and comprehension.
[60]	Concerns about faithfulness; users struggle to distinguish correct/incorrect explanations.	Plausible explanations increase trust but may mislead if incorrect, causing negative dominance.	Human-first interaction protocols significantly increased trust in AI predictions.
[61]	Faithful explanations needed for accurate risk factor identification and clinician trust.	Plausible explanations reveal meaningful factors and support clinical verification.	System usability scale (SUS) used; improved information gain, time-saving, and workflow integration noted.

[62]	May not fully reflect correct predictions; concern about following wrong advice.	Feature contribution explanations clearer; example-based less plausible alone; role-dependent trust.	Visual clarity improved ease of interpretation; prior CDSS experience influenced usability; variability in clinical protocols affected integration.
[63]	Limited faithfulness; debugging explanations misaligned with clinical decision processes.	Explanations unsatisfactory and not credible enough to boost explicit trust/confidence.	Mixed feedback: explanations increased trust but lowered ease of use; some found interface overwhelming and slow.
[64]	Partially faithful via interpretable features linked to patient records; complexity reduces faithfulness.	Plausible due to contextual info and visual hierarchy increasing acceptance.	Positive behavioral outcomes: smooth interaction, intuitive design, efficiency, and workflow integration reported.
[65]	Faithful SHAP mapping of feature contributions; highlights disagreements to support decision-making.	Aligns well with human understanding and clinical knowledge, increasing trust.	Usability reflected by time taken to complete validation tasks.
[66]	Explanations aligned with actual model decisions reflecting true behavior.	Matches known clinical risk factors, enhancing clinical meaningfulness and credibility.	Real-time feedback and clinician-driven design improved usability and workflow fit.

[67]	Accurate ClusteredSHAP explanations reflect the model decision process.	Clear, clinically meaningful insights enhance understanding and trust.	No usability details reported.
[68]	Unclear faithfulness; users struggled to grasp explanations.	Low plausibility; hard to understand, reducing trust and slowing decisions.	Behavioral intention, trust, effort expectancy, and facilitating conditions assessed; explanations sometimes complex.
[69]	Some skepticism due to subjective data representation gaps.	Mixed plausibility; sometimes incomplete or overly negative, lowering acceptance.	Preference for simplified explanations due to ease of use and satisfaction; some explanation types negatively impacted satisfaction.
[70]	Closely related to patient data, improving trust and alignment with real data patterns.	Visual and example-based explanations perceived as believable and actionable.	High intuitiveness, efficiency, and integration into clinical screening workflows; positive user satisfaction.
[71]	Handcrafted explanations may reduce faithfulness as they might not reflect true AI logic.	Designed to be clear and intuitive, increasing plausibility despite lower faithfulness.	Trust measured via scales; behavioral compliance with system advice tracked; issues with inappropriate reliance identified.

Appendix C

Table C.1 : User-Centred Approaches and Experimental Frameworks in Healthcare Explainability Studies

Study Number	Study design	User-Centered Approaches	Experimental Frameworks
[50]	Mixed Methods	Survey-based usability study, Clinician feedback	LIME (Local Interpretable Model-agnostic Explanations), ROC, F1, precision, recall, Jaccard similarity for performance evaluation
[51]	Mixed Methods	-	SHAP, Integrated Gradients, LIME, Faithfulness, AUROC, concept scores for explanation validation and comparison
[52]	Mixed Methods	Feedback from 7 medical professionals and 5 developers, Clinician assessment of explanation usefulness	Grad-CAM (Gradient-weighted Class Activation Mapping) for visualisation of model decisions, Cross-validation, Sensitivity, specificity, accuracy
[53]	Quantitative descriptive	-	SHAP, LIME, AraucanaXAI for evaluation based on identity, fidelity, separability, time. Performance compared on mortality prediction accuracy for ALS patients.
[54]	Qualitative Study	Think-aloud protocol, oncologists' evaluation of lung cancer relapse prediction explanations	No specific XAI methods mentioned but focuses on knowledge graph-based predictions and interpretability for oncologists' feedback.

[55]	Mixed Methods	Evaluation of Global Feature Importance and Local Feature Importance for enhancing trust in healthcare AI predictions	Random Forest model; Global Feature Importance (GFI), Local Feature Importance (LFI), evaluation of User Trust and Reliance (UTR) through Likert Scale.
[56]	Mixed Methods	Evaluation of XAI-driven user interfaces in neurological diagnosis (MCI/AD), focus on comprehension and trust	Brain connectivity analysis (ADNI dataset); uses Global Tornado Plot, Brain Connectivity Plot, and Counterfactual Plot for evaluation of visualisation and trust.
[57]	Mixed Methods	Human-AI collaboration, XAI impact on trust calibration	Quantitative: Repeated One-way ANOVA, Cognitive-based trust scale; Qualitative: Semi-structured interviews, content analysis
[58]	Mixed Methods	Human-centered design approach (DoReMi), user requirements for explainability	Quantitative: Median importance ratings, Likert scale; Qualitative: Open coding of interviews
[59]	Mixed Methods	Human-centered design, AI model interpretability	Quantitative: AUC, precision, recall, F1 score; Qualitative: Surveys, Likert scales, visualisation tools (LIME, SHAP)

[60]	Mixed Methods	User expertise, trust calibration, technology dominance in decision-making	Statistical analysis (Spearman ϱ), violin plots, scatter plots, and correlation analysis; AI system accuracy 70%
[61]	Mixed Methods	User-centered design focused on the needs of clinicians for explainability in XAI for HOB risk prediction. Used clinician feedback to guide the design of explainable AI systems.	The study involved 9 clinicians with expertise in microbiology, infectiology, and internal medicine. The data collection included interviews and questionnaires, focusing on clinicians' needs for explainability. The study developed a step-by-step implementation framework for XAI in healthcare.
[62]	Mixed Methods	Investigates the impact of two XAI methods (feature contribution vs. example-based) on healthcare practitioners' decision-making and advice-taking. Focus on clinicians'	Used a survey method with randomized case assignments. Data was gathered from healthcare practitioners, including obstetricians, midwives, and dietitians. The study utilized statistical tests to measure advice-taking and clinician preferences for XAI methods (e.g., WOA metric, preferences for explanation methods).

		preferences for XAI methods and how these methods affect their advice-taking behavior.	
[63]	Mixed Methods	The study investigates the impact of AI explanations on healthcare providers' decision-making, trust, and behavioral intention to use AI-based decision support systems. Focus on how explanations influence trust and confidence in algorithmic predictions.	Participants were healthcare providers, including doctors, nurses, and paramedics. Data collection was through an online experiment on the Prolific platform. The study used quantitative measures such as the Weight of Advice (WOA) and Behavioral Intention (BI), and qualitative feedback.
[64]	Qualitative Study	VBridge is a user-centered visual analytics tool designed to integrate machine learning explanations	Involved 6 clinicians from ZJUCH with an average of 17 years of experience. Data collection involved clinical data from pediatric cardiac surgery patients. Clinicians evaluated the VBridge system's effectiveness in improving trust in ML predictions and decision-making in a clinical setting.

		into clinicians' decision-making workflows. Focus on improving clinicians' understanding and trust in ML predictions for predicting surgical complications.	
[65]	Mixed Methods	Users reported higher levels of confidence, trust, and interpretability when using the XAI explanation module. It helped improve clinical decision-making efficiency by allowing medical professionals to quickly validate seizure predictions.	Sensitivity, specificity, precision metrics to evaluate models. Usability evaluation included validation time, interpretability, trust, and confidence using a Likert scale.

[66]	Mixed Methods	Feedback from expert clinicians assessed the tool's effectiveness in providing understandable and actionable explanations for diagnosing Age-related Macular Degeneration (AMD).	Accuracy, sensitivity, specificity, precision, F-score used to evaluate model performance. Evaluation of explanations based on expert clinician feedback (goodness, satisfaction, applicability).
[67]	Qualitative Study	Feedback from 30 medical experts was collected through a customized questionnaire to assess usability and explainability of the SHAP method optimizations.	Computational speed comparisons between different explainers (ClusteredSHAP, GradientExplainer, etc.). F1 score of original model performance. Explanation Usability Scores (EUS) based on aspects like understandability, accuracy, and trustworthiness.
[68]	Mixed Methods	Prototyping, Feedback collection from healthcare providers, Trust and satisfaction assessment	Empirical methods in interaction design, Interaction feedback, User study with healthcare providers (doctors, nurses, paramedics)

[69]	Qualitative Study	Focus groups, User feedback through discussions, Interaction with different explanation types	Thematic Analysis, User preference, and feedback on explanation types (Feature Importance and Counterfactual)
[70]	Mixed Methods	Focus groups, Evaluation of visual directive dashboard, Feedback on explanation types	Thematic Analysis, Statistical analysis, Success rate analysis, Task-based performance evaluation, User testing with healthcare providers and patients
[71]	Mixed Methods	Interaction with decision support systems (DSS), Perception of different explainable AI methods, Feedback from clinicians and general population	Descriptive statistics, ANOVA, Multiple linear regression, Logistic regression, Post-hoc analysis

Appendix D

Table D. 1: Assessment of Clinician and Patient Trustworthiness in XAI Explanations: Qualitative and Quantitative Methods

Framework Dimension	Mapped Studies	Trustworthiness Evaluation Metrics	Measurement Type

Perceived Trustworthiness	[53], [55], [57], [59], [60], [66],[67],[68], [69],	Clarity, reliability, explanation completeness, user trust perception, Confidence in model's reasoning, Trust in model predictions, Trust in system via visualisations	Qualitative: Participant feedback, open-ended responses, Subjective assessments, Observations of participant behaviour; Quantitative: 5-point Likert scales, binary ratings, verbal scales
Cognitive Understanding	[57], [59], [60], [63], [64], [67],[70],	Explanation clarity, decision path transparency, perceived technical competence, normalized Avg.EUS, Understanding of model's reasoning, Comprehension of visual explanations	Qualitative: Think-aloud protocols, open-ended feedback, Task behaviour, justifications, clinicians' observations Quantitative: Likert scales, Avg.EUS scores, ANOVA and Friedman test
Emotional & Behavioural Trust	[53], [55], [58], [65], [66], [71]	Confidence, reliance, behavioural intention, trust in predictions	Qualitative: Interview feedback, user justifications Quantitative: 5-/11-point Likert scales, Weight of Advice (WOA), BI metrics
Temporal Trust Dynamics	[63]	Initial vs. final trust level, trust evolution with exposure	Qualitative: Open-ended trust reasoning Quantitative: Likert scale comparisons over time, effect sizes, adjusted p-values
Usability & Integration	[60], [65], [66], [70]	Human-Computer Trust (HCT) scale, usability, likelihood of CDSS integration	Qualitative: Feedback on ease-of-use, system fit Quantitative: Likert ratings, SUS items, Tukey HSD, usability scores
Ethical & Regulatory Trust	[66]	Predictability, reliability, user attitude toward robotic systems (Negative Attitude Toward Robots Scale)	Qualitative: Thematic analysis of concerns Quantitative: 5-point Likert scales, NARS (score: 12–84)
Patient-Centered Trust	[59], [64], [69]	Trust based on how AI supports patient care decisions, understandability for clinical application	Qualitative: Interviews, open-ended feedback from clinicians Quantitative: Likert/Verbal ratings, NARS scale

References:

1. Walczak, S. (2020). The role of artificial intelligence in clinical decision support systems and a classification framework. In *Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications* (pp. 390-409). IGI Global Scientific Publishing.

2. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... & Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, *24*(9), 1342-1350.

3. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019. Oct 1;1(6):e271–97. doi: 10.1016/S2589-7500(19)30123-2

4. Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., ... & Z-Inspection Initiative. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016.

5. Hulsen, T. (2023). Explainable artificial intelligence (XAI): concepts and challenges in healthcare. *AI*, *4*(3), 652-666.
6. Mienye, I. D., Obaido, G., Jere, N., Mienye, E., Aruleba, K., Emmanuel, I. D., & Ogbuokiri, B. (2024). A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Informatics in Medicine Unlocked*, 101587.

7. Habibullah, K. M. (2024, June). Explainable AI: A Diverse Stakeholder Perspective. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)* (pp. 494-495). IEEE.
8. Ferrario, A., & Loi, M. (2022, June). How explainability contributes to trustworthiness in AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency(pp. 1457-1466).
9. Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. arXiv preprint arXiv:2004.03685.
10. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1-42.

11. Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI. *Applied Sciences*, *14*(23), 11288.
12. Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (xai). *IEEe Access*, *11*, 78994-79015.
13. Xing, X., Wu, H., Wang, L., Stenson, I., Yong, M., Del Ser, J., ... & Yang, G. (2024). Non-imaging medical data synthesis for trustworthy AI: A comprehensive survey. ACM Computing Surveys, 56(7), 1-35.
14. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1), 45-74.
15. Di Martino, F., & Delmastro, F. (2023). Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artificial Intelligence Review, 56(6), 5261-5315.
16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.
17. Caterson, J., Lewin, A., & Williamson, E. (2024). The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review. Digital health, 10, 20552076241272657.
18. Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., & Díaz-Rodríguez, N. (2021). Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950.
19. Adeniran, A. A., Onebunne, A. P., & William, P. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. World J. Adv. Res. Rev, 23, 2647-2658.
20. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthiness Artificial Intelligence. Information fusion, 99, 101805.
21. Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making*, *20*, 1-16.
22. Oberste, L., & Heinzl, A. (2022). User-centric explainability in healthcare: a knowledge-level perspective of informed machine learning. *IEEE Transactions on Artificial Intelligence*, *4*(4), 840-857.
23. Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In Computer science on-line conference (pp. 15-25). Cham: Springer International Publishing.

24. Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI systems evaluation: a review of human and computer-centred methods. Applied Sciences, 12(19), 9423.
25. Löfström, H., Hammar, K., & Johansson, U. (2022, May). A meta survey of quality evaluation criteria in explanation methods. In International Conference on Advanced Information Systems Engineering (pp. 55-63). Cham: Springer International Publishing.
26. Nayebi, A., Tipirneni, S., Foreman, B., Reddy, C. K., & Subbian, V. (2023, April). An empirical comparison of explainable artificial intelligence methods for clinical data: a case study on traumatic brain injury. In AMIA annual symposium proceedings (Vol. 2022, p. 815).
27. Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., ... & Kasneci, E. (2023). Towards human-centred explainable ai: A survey of user studies for model explanations. IEEE transactions on pattern analysis and machine intelligence, 46(4), 2104-2122.
28. Pawlicki, M., Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., & Choraś, M. (2024). Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing, 602, 128282.
29. Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine*, *166*, 107555.

30. Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer methods and programs in biomedicine*, *226*, 107161.

31. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., ... & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370.

32. Bharati, S., Mondal, M. R. H., & Podder, P. (2023). A review on explainable artificial intelligence for healthcare: Why, how, and when?. *IEEE Transactions on Artificial Intelligence*.

33. Al-Ansari, N., Al-Thani, D., & Al-Mansoori, R. S. (2024). User‐Centred Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Human Behaviour and Emerging Technologies*, *2024*(1), 4628855.

34. Aziz, N. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A., & Rashwan, W. (2024). Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems. *medRxiv*, 2024-08.

35. Prentzas, N., Kakas, A., & Pattichis, C. S. (2023). Explainable AI applications in the medical domain: A systematic review. *arXiv preprint arXiv:2308.05411*.

36. Nzenwata, U. J., OO, I., Tai-Ojuolape, E. O., Aderogba, T. A., Durodola, O. F., Kesinro, P. O., ... & Adesuyan, M. A. (2024). Explainable AI: A Systematic Literature Review Focusing on Healthcare. *Journal of Computer Sciences and Applications*, *12*(1), 10-16.

37. Shafik, W., Hidayatullah, A. F., Kalinaki, K., Gul, H., Zakari, R. Y., & Tufail, A. (2024). A Systematic Literature Review on Transparency and Interpretability of AI models in Healthcare: Taxonomies, Tools, Techniques, Datasets, OpenResearch Challenges, and Future Trends.

38. Eke, C. I., & Shuib, L. (2024). The role of explainability and transparency in fostering trustworthiness in AI healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Computing and Applications*, 1-36.

39. Hamida, S. U., Chowdhury, M. J. M., Chakraborty, N. R., Biswas, K., & Sami, S. K. (2024). Exploring the landscape of explainable artificial intelligence (XAI): A systematic review of techniques and applications. *Big Data and Cognitive Computing*, *8*(11), 149.

40. Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, *12*(3), 1353.

41. Brdnik, S., & Šumak, B. (2024, May). Current trends, challenges and techniques in XAI field; A tertiary study of XAI research. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 2032-2038). IEEE.

42. Kolluru, V., Nuthakki, Y., Mungara, S., Koganti, S., Chintakunta, A. N., & Telaganeni, C. S. (2023). Healthcare Through AI: Integrating Deep Learning, Federated Learning, and XAI for Disease Management. *International Journal of Soft Computing and Engineering (IJSCE)*, *13*, 21-30.
43. S. Keele et al., “Guidelines for performing systematic literature reviews in software engineering,” 2007.

44. *PRISMA 2020 statement — PRISMA statement*. (n.d.). PRISMA Statement. https://www.prisma-statement.org/prisma-2020
45. Khan, Khalid, S., ter Riet, Gerben., Glanville, Julia., Sowden, Amanda, J. and Kleijnen, Jo. (eds) Undertaking Systematic Review of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition), NHS Centre for Reviews and Dissemination, University of York, IBSN 1 900640 20 1, March 2001.
46. Cochrane Collaboration. Cochrane Reviewers' Handbook. Version 4.2.1. December 2003
47. Centre for Reviews and Dissemination. (2008). *Systematic reviews*. CRD, University of York. https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf
48. Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/efi-180221
49. Carver, J. C., Hassler, E., Hernandes, E., & Kraft, N. A. (2013, October). Identifying barriers to the systematic literature review process. In *2013 ACM/IEEE international symposium on empirical software engineering and measurement* (pp. 203-212). IEEE.
50. Hendawi, R., Li, J., & Roy, S. (2023). A mobile app that addresses interpretability challenges in machine learning–based diabetes predictions: survey-based user study. *JMIR Formative Research*, *7*(1), e50328.
51. Ghanvatkar, S., & Rajan, V. (2024). Evaluating Explanations From AI Algorithms for Clinical Decision-Making: A Social Science-Based Approach. *IEEE Journal of Biomedical and Health Informatics*.
52. Oprea, C., Grüne, M., Buglowski, M., Olivier, L., Orlikowsky, T., Kowalewski, S., ... & Stollenwerk, A. (2024). Evaluating the explainable ai method grad-cam for breath classification on newborn time series data. *IFAC-PapersOnLine*, *58*(24), 123-128.
53. Buonocore, T. M., Nicora, G., Dagliati, A., & Parimbelli, E. (2022). Evaluation of XAI on ALS 6-months mortality prediction. In *CLEF (Working Notes)* (pp. 1228-1235).
54. Anjara, S. G., Janik, A., Dunford-Stenger, A., Mc Kenzie, K., Collazo-Lorduy, A., Torrente, M., ... & Provencio, M. (2023). Examining explainable clinical decision support systems with think aloud protocols. *Plos one*, *18*(9), e0291443.
55. Polat Erdeniz, S., Veeranki, S., Schrempf, M., Jauk, S., Ngoc Trang Tran, T., Felfernig, A., ... & Leodolter, W. (2022, September). Explaining machine learning predictions of decision support systems in healthcare. In *Current Directions in Biomedical Engineering* (Vol. 8, No. 2, pp. 117-120). De Gruyter.
56. Lombardi, A., Marzo, S., Di Noia, T., Di Sciascio, E., & Ardito, C. (2024, June). Exploring the usability and trustworthiness of AI-driven user interfaces for neurological diagnosis. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 627-634).

57. Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trustworthiness calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*, 102941.
58. Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., & Van Den Bosch, K. (2021). Human-centred XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, *154*, 102684.
59. Dwiyanti, L., Nambo, H., & Hamid, N. (2024). Leveraging Explainable Artificial Intelligence (XAI) for Expert Interpretability in Predicting Rapid Kidney Enlargement Risks in Autosomal Dominant Polycystic Kidney Disease (ADPKD). *AI*, *5*(4), 2037-2065.
60. Cabitza, F., Campagner, A., Natali, C., Parimbelli, E., Ronzio, L., & Cameli, M. (2023). Painting the black box white: experimental findings from applying XAI to an ECG reading setting. *Machine Learning and Knowledge Extraction*, *5*(1), 269-286.
61. Hoogestraat, A. T., & Wulff, A. (2024). A Vision on User-Centred Implementation and Evaluation of Explainable AI for Predicting Hospital-Onset Bacteremia. *Studies in health technology and informatics*, *316*, 766-770.
62. Du, Y., Antoniadi, A. M., McNestry, C., McAuliffe, F. M., & Mooney, C. (2022). The role of XAI in advice-taking from a clinical decision support system: a comparative user study of feature contribution-based and example-based explanations. *Applied Sciences*, *12*(20), 10323.
63. Panigutti, C., Beretta, A., Giannotti, F., & Pedreschi, D. (2022, April). Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-9).
64. Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., ... & Veeramachaneni, K. (2021). Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on visualisation and Computer Graphics*, *28*(1), 378-388.
65. Raab, D., Theissler, A., & Spiliopoulou, M. (2023). XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Neural Computing and Applications*, *35*(14), 10051-10068.
66. Martínez-Villaseñor, L., Ponce, H., Martínez-Velasco, A., & Miralles-Pechuán, L. (2022, July). An explainable tool to support age-related macular degeneration diagnosis. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
67. Mo, B. Y., Nuannimnoi, S., Baskoro, A., Khan, A., Ariesta Dwi Pratiwi, J., & Huang, C. Y. (2023, December). ClusteredSHAP: Faster GradientExplainer based on K-means Clustering and Selections of Gradients in Explaining 12-Lead ECG Classification Model. In *Proceedings of the 13th International Conference on Advances in Information Technology* (pp. 1-8).

68. Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of human-centred, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1-35.
69. Szymanski, M., Vanden Abeele, V., & Verbert, K. (2024, May). Designing and evaluating explanations for a predictive health dashboard: A user-centred case study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-8).
70. Bhattacharya, A., Ooge, J., Stiglic, G., & Verbert, K. (2023, March). Directive explanations for monitoring the risk of diabetes onset: introducing directive data-centric explanations and combinations to support what-if explorations. In *Proceedings of the 28th international conference on intelligent user interfaces* (pp. 204-219).
71. Gombolay, G. Y., Silva, A., Schrum, M., Gopalan, N., Hallman‐Cooper, J., Dutt, M., & Gombolay, M. (2024). Effects of explainable artificial intelligence in neurology decision support. *Annals of clinical and translational neurology*, *11*(5), 1224-1235.