

**Explainable Artificial Intelligence (XAI) for Improving Trust and
Transparency in Healthcare Machine Learning Systems: A Literature Review**

Student Preparation

Under the supervision of

2026

Introduction

One of the most revolutionary technologies in healthcare today is artificial intelligence (AI), which supports clinical decision-making, patient monitoring, medical picture analysis, disease prediction, and diagnosis. In numerous healthcare applications, deep learning and machine learning models have demonstrated remarkable predictive performance. Despite these advancements, healthcare practitioners are frequently reluctant to depend on AI systems because many sophisticated models function as "black boxes," making predictions without offering clear justifications for their choices. Physicians must defend their diagnostic and treatment choices to patients, healthcare organizations, and regulatory bodies, which presents difficulties in clinical practice due to this lack of openness. As a result, Explainable Artificial Intelligence (XAI), which seeks to improve the transparency, interpretability, and reliability of AI models, has drawn more attention from researchers. Clinicians can confirm AI recommendations and boost confidence in AI-assisted decision-making by using XAI's explanations of prediction generation. Although there are still issues with measuring explanation quality and striking a balance between predictive accuracy and model interpretability, recent systematic evaluations show that XAI is becoming a crucial prerequisite for the safe application of AI in healthcare.

Literature Review

Paper 1

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)*.

- Research Objective

The primary objective of this systematic review was to examine how Explainable Artificial Intelligence has been applied in healthcare over the previous decade. The authors aimed to identify the major healthcare domains utilizing XAI, compare the explainability techniques employed, evaluate the datasets used, and identify research gaps requiring future investigation.

- Methodology

The researchers conducted a systematic literature review following PRISMA guidelines. Studies published between 2011 and 2022 were retrieved from major scientific databases. After applying inclusion and exclusion criteria, the selected studies were analyzed according to healthcare application, AI model, XAI technique, dataset characteristics, and performance evaluation.

- Key Findings

According to the review, XAI has been effectively used in clinical decision support, predictive medicine, medical imaging, electronic health records, and illness diagnosis. SHAP, LIME, Grad-CAM, and attention mechanisms were the most commonly employed explanation strategies. The scientists came to the conclusion that although deep learning models performed exceptionally well in terms of prediction, their adoption in the medical field is mostly dependent on clear justifications that physicians can comprehend and rely on.

- Critical Analysis

Strengths

- Comprehensive systematic review covering more than ten years of research.
- Well-organized classification of XAI techniques.
- Extensive discussion of healthcare applications.
- Clear identification of future research opportunities.

Weaknesses

- Limited discussion regarding clinical implementation.
- Few studies included real-world deployment.
- Most reviewed research emphasized technical performance rather than physician usability.

Paper 2

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). *Essential Properties and Explanation Effectiveness of Explainable Artificial Intelligence in Healthcare: A Systematic Review*.

- Research Objective

This study investigated the characteristics of effective explanations produced by healthcare AI systems. Rather than focusing solely on predictive performance, the authors examined whether explanations improve user trust, understanding, and decision-making.

- Methodology

The authors systematically searched PubMed and Embase databases. Six eligible studies met the strict inclusion criteria and were analyzed according to explanation quality, stakeholder needs, trust evaluation, and human-centered effectiveness measures.

- Key Findings

The review demonstrated that explanation effectiveness depends on several dimensions, including fidelity, plausibility, interpretability, and user satisfaction. The authors found no standardized framework for evaluating explanation quality, making comparison between studies difficult. They recommended developing internationally accepted evaluation standards for healthcare XAI.

- Critical Analysis

Strengths

- Strong emphasis on human-centered AI.
- Comprehensive evaluation framework.
- Focus on clinician trust.

Weaknesses

- Small number of eligible studies.
- Limited empirical validation.
- Narrow inclusion criteria reduced generalizability.

Paper 3

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). *Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review*.

- Research Objective

This study's main goal was to thoroughly examine the use of Explainable Artificial Intelligence (XAI) models that make use of actual Electronic Health Record (EHR) data. The authors aimed to find out how explainability approaches have been integrated into machine learning models created from EHR data, identify popular explanation techniques, look at their applications in healthcare, and ascertain the present obstacles preventing their clinical use. The study also sought to offer suggestions for further investigation that would improve the dependability and transparency of AI systems in medical settings.

- Methodology

In accordance with PRISMA-ScR criteria, the researchers carried out a systematic scoping review. Studies published up to 2020 were found by searching a number of electronic sources, including Web of Science, IEEE Xplore, PubMed, and the ACM Digital Library. The chosen studies were examined based on healthcare application, machine learning algorithm, explainability technique, data characteristics, assessment metrics, and reported clinical outcomes following the application of predetermined inclusion and exclusion criteria. Instead of statistically summarizing findings, the authors were able to identify knowledge gaps and present a comprehensive overview of current research thanks to the scoping review process.

- Key Findings

According to the review, one of the most useful sources of medical data for creating predictive machine learning models is electronic health records. Deep neural networks, random forests, gradient boosting machines, and tree-based techniques were used in the majority of the examined studies. SHAP and LIME were the most commonly used explanation strategies because of their capacity to pinpoint significant clinical variables that influence predictions.

The authors discovered that explainable models were frequently employed to forecast the course of chronic diseases, hospital readmission, mortality, disease diagnosis, and intensive care unit

outcomes. Nevertheless, many research lacked external validation and practical clinical application, despite promising predictive performance. The evaluation stressed that before widespread adoption can take place, explanations must be both clinically significant and comprehensible to doctors. Transparency alone is not enough.

- Critical Analysis

Strengths

- Comprehensive review focusing specifically on real-world Electronic Health Record data.
- Included a wide range of healthcare prediction tasks.
- Clearly categorized explainability techniques and machine learning algorithms.
- Identified practical challenges associated with clinical implementation.
- Followed an internationally recognized systematic review methodology.

Weaknesses

- Most reviewed studies were retrospective rather than prospective.
- Limited evidence regarding physician acceptance of explanations.
- Few studies performed external validation using independent datasets.
- Did not quantitatively compare explanation quality across different XAI techniques.

Paper 4

Allgaier, J., Mulansky, L., Draelos, R., & Pryss, R. (2023). *How Does the Model Make Predictions? A Systematic Literature Review on the Explainability Power of Machine Learning in Healthcare.*

- Research Objective

Examining explainability's integration into healthcare machine learning models was the aim of this comprehensive literature study. The study's main objectives were to determine the most popular explainability techniques, analyze the healthcare fields where XAI has been successfully applied, and determine whether the explanations currently in use are enough instructive for medical practitioners. Additionally, the authors aimed to investigate if current research sufficiently analyzes the utility of explanations from a clinical standpoint.

- Methodology

The published research articles that were obtained from significant scientific databases were systematically reviewed by the writers. Studies were categorized by application field, machine learning model, explanation technique, assessment strategy, and target consumers after being screened in accordance with predetermined eligibility criteria. Differentiating between post hoc explanation strategies and models that are intrinsically interpretable received special attention.

- Key Findings

The analysis showed that explainable machine learning is most commonly used in medical imaging, especially in radiology, pathology, and cancer diagnosis. While structured clinical datasets mostly used SHAP and LIME, deep learning models often used visualization techniques like Grad-CAM and saliency maps.

The authors discovered that only a small number of research explicitly assessed explanations with healthcare professionals, despite the fact that explanation strategies increased transparency. As a result, many explanation techniques continue to be technically impressive but lack clinical validation. In order to increase the explanations' practical utility, the evaluation suggested include clinicians at every stage of the AI development process.

- Critical Analysis

Strengths

- Comprehensive comparison of multiple explainability techniques.
- Strong focus on healthcare applications.

- Clear distinction between intrinsic and post hoc interpretability.
- Identified important gaps in human-centered evaluation.

Weaknesses

- Limited discussion of ethical and legal considerations.
- Most reviewed studies relied on retrospective datasets.
- Few studies included clinician feedback during model evaluation.
- Lack of standardized comparison metrics across studies.

Paper 5

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). *Survey of Explainable AI Techniques in Healthcare*.

- Research Objective

The goal of this survey was to give a thorough overview of Explainable Artificial Intelligence methods that are currently being applied in the medical field. The authors specifically looked at the development of XAI techniques, classified various explanation strategies, assessed their use in various medical specialties, and talked about the obstacles to wider clinical implementation. Future research avenues for enhancing transparency, equity, and dependability in healthcare AI systems were also examined in the poll.

- Methodology

A narrative survey of recently released research on Explainable Artificial Intelligence was carried out by the researchers. The explanation method, machine learning model, healthcare application, evaluation approach, and future research direction were used to categorize the articles. In contrast to systematic reviews, the survey did not limit inclusion to particular databases or study types; instead, it synthesized wider trends within the field.

- Key Findings

The poll found that because healthcare decisions need to be transparent and accountable, Explainable Artificial Intelligence has grown in significance. The most popular explainability techniques were found to include SHAP, LIME, Grad-CAM, attention processes, and counterfactual explanations.

Explainable models boost patient trust, promote regulatory compliance, and boost professional confidence, according to the authors. Nevertheless, they stressed that many current explanation approaches remain difficult for non-technical healthcare professionals to interpret. The survey suggested creating explanation techniques that strike a compromise between usability, simplicity, and prediction accuracy.

- Critical Analysis

Strengths

- Broad overview covering numerous XAI techniques.
- Excellent explanation of different categories of interpretability.
- Discussed technical, ethical, and practical challenges.
- Included recent developments in deep learning explainability.

Weaknesses

- Narrative survey rather than systematic review.
- No quantitative comparison between explanation methods.
- Limited discussion regarding implementation costs.
- Some conclusions were based primarily on theoretical discussions.

Paper 6

A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges (2024).

- Research Objective

By examining current developments, new uses, and unsolved research issues, this survey aimed to present an updated overview of Explainable Artificial Intelligence in healthcare. The study investigated future prospects for incorporating XAI into clinical decision support systems and looked at explainability's role in reliable AI systems.

- Methodology

The authors reviewed a wide range of recent research on Explainable Artificial Intelligence in Healthcare. The healthcare domain, explanation strategy, AI model, evaluation methodology, clinical usefulness, ethical considerations, and implementation obstacles were taken into consideration when analyzing the evaluated papers.

- Key Findings

According to the report, Explainable AI has quickly spread to a wide range of healthcare applications, such as medical imaging, precision medicine, disease diagnosis, critical care monitoring, and individualized treatment planning. The authors stressed that explainability boosts accountability in AI-assisted healthcare, accelerates regulatory approval, and enhances clinician trust.

The survey revealed a number of outstanding challenges despite notable advancements, such as uneven evaluation techniques, inadequate external validation, low physician engagement, computational complexity, and the lack of widely recognized explainability standards. The authors came to the conclusion that human-centered design and interdisciplinary cooperation between computer scientists, physicians, and healthcare regulators should be given top priority in future healthcare AI systems.

- **Critical Analysis**

Strengths

- Covers the latest developments in healthcare Explainable Artificial Intelligence.
- Comprehensive discussion of technical, ethical, and regulatory challenges.
- Strong emphasis on future research directions.
- Includes multiple healthcare applications and explanation methods.
- Highlights the importance of interdisciplinary collaboration.

Weaknesses

- Survey-based methodology does not statistically synthesize findings.
- Limited empirical comparison of competing explanation techniques.
- Some emerging methods lack sufficient real-world clinical evidence.
- Many reviewed studies were conducted in controlled research environments rather than routine healthcare settings.

Connections and Conflicts Between the Reviewed Papers

Explainability is crucial for boosting clinician trust and promoting the use of AI in healthcare, according to the examined publications. Transparency is a requirement for incorporating AI into standard clinical operations, according to all six research. They also agree that post hoc explanation approaches such as SHAP and LIME are extensively employed but still require rigorous validation before being relied upon in high-stakes medical choices.

There are some distinctions, though. While Loh et al. (2022) emphasize increasing XAI applications across healthcare disciplines, Jung et al. (2023) suggest that enhancing explanation quality and developing consistent evaluation criteria should be emphasized before widespread implementation. The conflict between preserving interpretability and optimizing prediction accuracy is also highlighted in other survey articles, suggesting that more accurate deep learning models are frequently less explicable.

Summary Table

Year	Author(s)	Methodology	Key Findings	Weaknesses
2022	Loh et al.	Systematic literature review	XAI improves transparency and clinician trust across multiple healthcare applications	Limited discussion of real-world implementation
2023	Jung et al.	Systematic review	No standardized framework exists for evaluating explanation effectiveness	Small sample of eligible studies
2020	Payrovnaziri et al.	Scoping review	Electronic health records are a major application area for XAI	Limited evaluation of deployment outcomes

2023	Allgaier et al.	Systematic review	Most healthcare XAI research focuses on medical imaging	Limited clinical validation
2023	Chaddad et al.	Survey	SHAP and LIME are the dominant explanation methods	Lack of standardized benchmarking
2024	Recent healthcare XAI survey	Comprehensive review	Future research should emphasize trustworthy and human-centered AI	Need for international standards

Overall Synthesis

Explainable AI has developed from a specialist research area into a crucial prerequisite for reliable healthcare AI, as the examined literature shows. Even while deep learning and machine learning models continue to attain excellent predicted accuracy, their general adoption depends on offering explanations that medical professionals can comprehend and assess. Transparency, accountability, and user trust are the main elements determining the effective application of AI in healthcare, according to all six studies. Important research gaps are also identified in the literature, such as the lack of defined evaluation measures, the paucity of real-world clinical validation, and the inadequate incorporation of clinician feedback into the design of XAI systems. Therefore, human-centered evaluation frameworks, standardized benchmarking techniques, and prospective clinical studies that concurrently evaluate explanation quality and prediction performance should be given top priority in future research.

References

- Allgaier, J., Mulansky, L., Draelos, R. L., & Pryss, R. (2023). *How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artificial Intelligence in Medicine*, 143, 102616. <https://doi.org/10.1016/j.artmed.2023.102616>
- Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). *Survey of explainable AI techniques in healthcare. Sensors*, 23(2), 634. <https://doi.org/10.3390/s23020634>
- Jung, J., Lee, H., Jung, H., & Kim, H. (2023). *Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon*, 9(5), e16110. <https://doi.org/10.1016/j.heliyon.2023.e16110>
- 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. <https://doi.org/10.1016/j.cmpb.2022.107161>
- Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). *Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. Journal of the American Medical Informatics Association*, 27(7), 1173–1185. <https://doi.org/10.1093/jamia/ocaa053>
- A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges. (2024). *Intelligent Systems with Applications*. <https://doi.org/10.1016/j.iswa.2024.200419>