Journal Of Artificial Intelligence In Medical Sciences

Vol.01-02-2026

EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE SYSTEMS FOR ENHANCING CLINICAL DECISION-MAKING AND PATIENT SAFETY IN HEALTHCARE

Sana Afzal

BS, Department of Computer Science, The Cosmic Degree College (For Women) ,University of Okara

ABSTRACT

Artificial intelligence (AI) is increasingly influencing healthcare through applications in disease diagnosis, clinical risk prediction, treatment recommendations, and patient monitoring. However, many advanced AI systems operate as complex black-box models, making it difficult for healthcare professionals to understand the reasoning behind their predictions and recommendations. This lack of transparency can create challenges involving clinical trust, accountability, diagnostic reliability, and patient safety. Explainable artificial intelligence (XAI) has emerged as an important approach for addressing these limitations by providing understandable explanations of AI-generated outputs. Techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM), feature importance analysis, and attention-based visualization can help clinicians examine the factors influencing model predictions. This research article evaluates the contribution of explainable AI systems to clinical decision-making and patient safety, with particular attention to diagnostic accuracy, treatment planning, risk assessment, transparency, and healthcare professional trust. It also examines the technical, ethical, and organizational challenges associated with implementing explainable AI in clinical environments. The study adopts a qualitative descriptive review methodology based on relevant peer-reviewed literature published between 2025 and 2026. The reviewed evidence suggests that explainability can improve the interpretability and usability of AI-supported clinical decision systems when explanations are accurate, understandable, and relevant to clinical workflows. Nevertheless, explanation methods do not automatically guarantee correct predictions or safer clinical decisions. Limitations involving explanation fidelity, external validation, algorithmic bias, workflow integration, and overreliance on automated recommendations remain significant concerns. The article concludes that explainable AI should be developed through human-centered evaluation, rigorous clinical validation, and appropriate professional oversight to support trustworthy and safer healthcare decision-making.

KEYWORDS

Explainable Artificial Intelligence, Clinical Decision-Making, Patient Safety, Clinical Decision Support Systems, Healthcare Transparency, SHAP, LIME, AI Interpretability.

INTRODUCTION

Artificial intelligence has become an important technological development in contemporary healthcare, offering new possibilities for improving disease diagnosis, clinical prediction, treatment planning, and patient management. Machine learning and deep learning models can analyze large volumes of medical information, identify complex patterns, and generate predictions that may assist healthcare professionals in making clinical decisions. These applications include medical image interpretation, disease risk assessment, patient deterioration

prediction, clinical documentation analysis, and personalized treatment recommendations. Despite their growing capabilities, many advanced AI systems rely on complex computational architectures that make their internal decision-making processes difficult to understand. This limitation has created concerns regarding transparency, reliability, professional accountability, and patient safety. Explainable artificial intelligence (XAI) has therefore emerged as an important research area focused on making AI-generated predictions more understandable and clinically interpretable (Abbas et al., 2025; Salimparsa et al., 2025).

Clinical decision-making involves the interpretation of multiple types of information, including patient history, physical examination findings, laboratory investigations, medical imaging, and established clinical guidelines. Healthcare professionals must evaluate this evidence before selecting diagnostic or therapeutic interventions. AI-supported clinical decision support systems can assist this process by identifying relevant patterns and generating predictions or recommendations. However, when an AI system provides a recommendation without explaining the factors responsible for its output, clinicians may find it difficult to assess whether the result is clinically reasonable. This issue becomes particularly important when AI predictions influence decisions involving medication selection, diagnostic procedures, emergency treatment, or intensive care management. Salimparsa et al. (2025) explain that the practical usefulness of explainable AI depends not only on the technical quality of explanations but also on their alignment with clinicians' information needs and everyday healthcare workflows.Explainable AI refers to a collection of methods designed to make the behavior, predictions, and decision processes of artificial intelligence systems more understandable to human users. It can provide information about which patient characteristics, clinical variables, image regions, or other data features contributed to a particular prediction. Some methods explain individual predictions, while others provide an overall understanding of model behavior. SHAP estimates the contribution of individual features to model predictions, whereas LIME generates locally interpretable approximations of complex models. Grad-CAM highlights image regions that influence deep learning predictions, particularly in medical imaging applications. These techniques can help healthcare professionals examine the evidence behind AI-generated outputs, although their reliability and usefulness depend on the underlying model and the explanation method being applied (Abbas et al., 2025; Esfahani et al., 2025)Transparency is particularly important when AI systems are used in high-risk clinical environments. A prediction involving the possibility of sepsis, cardiovascular deterioration, neurological complications, or serious medication-related harm may influence urgent clinical decisions. In such circumstances, healthcare professionals need to understand the factors associated with an AI-generated risk assessment and determine whether those factors are consistent with the patient's clinical condition. Explainable systems may provide information about influential variables, identify relevant clinical indicators, and help users recognize potential inconsistencies in automated recommendations. However, an explanation should not be treated as independent proof that a prediction is correct. An AI model may generate an inaccurate prediction and still provide an

apparently convincing explanation. Consequently, explanation quality must be evaluated separately from predictive performance (Bauer & Michalowski, 2025; Zhang et al., 2026).

The relationship between explainable AI and patient safety extends beyond improving the visibility of algorithmic decisions. Patient safety requires healthcare professionals to identify potential errors, recognize clinical risks, and avoid inappropriate treatment decisions. An explainable decision support system may help clinicians identify which patient characteristics influenced a recommendation and whether the system relied on clinically relevant information. For example, in a risk prediction system, an explanation may indicate that abnormal laboratory measurements, age-related factors, or specific physiological observations contributed to a high-risk classification. Such information can support further clinical assessment and encourage professionals to verify the recommendation against available evidence. Nevertheless, patient safety depends on the complete clinical process, including data quality, model validation, professional judgment, and appropriate monitoring rather than explainability alone (Abbas et al., 2025).Healthcare professional trust is another important factor influencing the successful adoption of AI-supported clinical systems. Physicians, nurses, pharmacists, and other healthcare workers may be reluctant to rely on systems that produce recommendations without understandable justification. Tun et al. (2025) examined trust in AI-based clinical decision support systems among healthcare workers and identified transparency, training, usability, clinical reliability, validation, ethical considerations, human-centered design, and professional control as important factors influencing trust. Their findings indicate that explainability must be supported by broader organizational and professional measures. Healthcare workers need appropriate training and opportunities to evaluate system performance, while AI tools should preserve their ability to exercise independent clinical judgment.Explainability also plays an important role in the evaluation of diagnostic and predictive models. Conventional machine learning evaluation frequently focuses on measures such as accuracy, sensitivity, specificity, precision, recall, and area under the receiver operating characteristic curve. Although these indicators provide useful information about predictive performance, they do not establish whether clinicians can understand or appropriately use a model's recommendations. Bauer and Michalowski (2025) emphasize the importance of human-centered evaluation and identify interpretability, fidelity, and clinical value as important dimensions of explainability assessment. This perspective suggests that clinical AI systems should be evaluated through both technical measurements and assessments involving healthcare professionals who use the systems in practice.

Despite its potential, explainable AI faces several technical and practical challenges. Some explanation methods may provide unstable results when input data change slightly, while others may generate visualizations that are difficult for clinicians to interpret. Complex explanations can also increase cognitive workload, particularly in emergency departments and other time-sensitive environments. Furthermore, an explanation may appear understandable without

accurately representing the actual behavior of the underlying model. Roychowdhury et al. (2025) highlight the importance of balancing technical fidelity with human interpretability and identify limited patient involvement, usability concerns, and insufficient clinical integration as continuing research gaps. These challenges demonstrate that explanation methods must be evaluated according to their accuracy, stability, relevance, and practical usefulness.Ethical and regulatory considerations are equally important in the development of explainable AI systems. Healthcare algorithms are trained using patient information that may contain sensitive medical and personal details. Data protection, informed governance, cybersecurity, and responsible information management are therefore necessary throughout model development and deployment. Algorithmic bias is another significant concern because training datasets may not adequately represent all patient populations. An AI system may perform differently across age groups, sexes, geographical populations, or healthcare institutions. Esfahani et al. (2025) emphasize the need for broader clinical validation, improved governance, and collaboration among technical researchers, healthcare professionals, and other stakeholders. Accordingly, explainable AI should be developed as part of a wider framework of trustworthy healthcare technology rather than treated as a purely technical feature.The growing importance of AI in healthcare makes it necessary to evaluate whether explainable systems can genuinely improve clinical decision-making and patient safety. The availability of explanation tools does not automatically establish their clinical value, and further research is required to determine whether they improve decision quality, professional understanding, and patient-related outcomes. Zhang et al. (2026) emphasize the importance of connecting technical explanation methods with clinical utility, workflow alignment, and stakeholder requirements. This research therefore examines explainable AI systems as a combination of technical methods, clinical decision-support tools, and human-centered healthcare technologies. It evaluates their potential benefits while identifying the limitations that must be addressed before their wider clinical adoption.

RESEARCH OBJECTIVES

1. To examine the role of explainable artificial intelligence systems in improving transparency and supporting clinical decision-making in healthcare.
2. To evaluate the contribution of explainability techniques to patient safety, clinical trust, and the reliability of AI-supported recommendations.
3. To identify the technical, ethical, and organizational challenges affecting the implementation of explainable AI systems in healthcare settings.

RESEARCH QUESTIONS

1. How do explainable artificial intelligence systems improve transparency and support clinical decision-making in healthcare?
2. To what extent can explainability techniques contribute to patient safety, healthcare

professional trust, and reliable clinical recommendations?
3. What technical, ethical, and organizational challenges influence the successful implementation of explainable AI systems in clinical practice?

LITERATURE REVIEW

Abbas et al. (2025) examined explainable artificial intelligence in clinical decision support systems through a systematic review and meta-analysis of 62 peer-reviewed studies published between 2018 and 2025. Their research investigated the application of XAI techniques across clinical areas, including radiology, oncology, neurology, and critical care. The authors identified visualization-based approaches, including Grad-CAM and attention mechanisms, as important techniques in imaging and sequential data applications. Their review also highlighted continuing limitations in explanation fidelity, methodological transparency, clinician trust, and real-world usability evaluation. The authors emphasized that clinical adoption requires longitudinal validation, participatory system design, and more consistent interpretability measures. This research demonstrates that explainability is an important component of clinical decision support, but its value must be established through rigorous technical and human-centered evaluation.Salimparsa et al. (2025) investigated explainable AI for clinical decision support systems by examining existing research, major methodological gaps, and opportunities for more effective implementation. Their review identifies the opacity of complex AI models as an important barrier to clinical adoption. The authors explain that technical explanations may fail to satisfy healthcare professionals when they are not aligned with clinical workflows or user requirements. They propose a user-centered framework involving appropriate XAI method selection, interface co-design, and iterative evaluation. Their research also emphasizes calibrated trust, suggesting that healthcare professionals should neither reject reliable AI recommendations unnecessarily nor accept automated outputs without critical assessment. This literature is particularly relevant to patient safety because it connects explanation design with the practical conditions under which clinical decisions are made.Tun et al. (2025) conducted a systematic review examining trust in AI-based clinical decision support systems among healthcare workers. Their study included 27 publications and identified several factors influencing professional acceptance of AI-supported decision systems. These factors included transparency, training, usability, clinical reliability, system validation, ethical considerations, human-centered design, and professional autonomy. The authors explain that healthcare workers may hesitate to use AI systems when their recommendations are difficult to understand or when the technology does not fit existing clinical workflows. Their findings also emphasize the importance of preserving healthcare professionals' control over final decisions. This research indicates that explainability contributes to trust when it is combined with reliable performance, appropriate training, and meaningful professional involvement.Bauer and Michalowski (2025) examined human-centered explainability evaluation in clinical decision-making through a critical review of the literature. Their research focused on how healthcare professionals evaluate explanations generated by XAI

methods for patient-specific clinical decisions. The review included 25 studies and identified substantial variation in the methods used to assess explainability. The authors highlighted three important dimensions of clinical explainability: interpretability, fidelity, and clinical value. Interpretability concerns whether users can understand an explanation, fidelity addresses whether the explanation accurately represents the model's behavior, and clinical value concerns whether the explanation supports meaningful clinical decisions. Their findings demonstrate that technical explanations should not be evaluated independently of the healthcare professionals expected to use them. This work supports the need for human-centered assessment when determining whether XAI systems can contribute to safer clinical practice.

Roychowdhury et al. (2025) investigated the evaluation of explanation performance in clinical decision support systems using non-imaging medical data. Their systematic literature review examined explanation quality, clinical decision-making, user trust, and usability. The authors identified applications across intensive care, oncology, neurology, and other clinical decision-support environments. Their review found that studies commonly combine technical evaluation measures, such as predictive accuracy and explanation fidelity, with human-centered assessments involving trust and usability. However, patient participation, explanation usability, and clinical integration remain insufficiently addressed in much of the available research. The authors emphasize that effective explainability requires a balance between technical faithfulness and human interpretability. Their findings are important because they demonstrate that explanations must be both technically meaningful and understandable to healthcare professionals.Esfahani et al. (2025) reviewed the application of explainable artificial intelligence in the prediction, diagnosis, treatment, and management of chronic diseases. Their research examined the use of different explanation methods across chronic disease applications and identified SHAP, LIME, and Grad-CAM among the commonly used techniques. The review indicated that SHAP was frequently applied to structured clinical information, while Grad-CAM was particularly relevant to medical imaging applications. The authors also identified limitations involving complex medical data, insufficient data volume, and the need for stronger real-world clinical validation. Their findings demonstrate that explainability methods must be selected according to the type of medical data and the clinical purpose of the AI system. The study further supports the need to extend explainable AI research beyond prediction toward treatment planning and long-term disease management.

Zhang et al. (2026) presented a comprehensive review of explainable artificial intelligence methods, evaluation approaches, and clinical integration. Their research emphasizes that the development of explanation techniques alone is insufficient to ensure successful healthcare implementation. The authors propose an approach that considers the properties of explanations, their dependence on the underlying model, and the scope of the clinical task. They also emphasize the importance of combining technical evaluation, including explanation fidelity and robustness, with clinical utility measures such as workflow alignment and clinician confidence.

Their review identifies the different requirements of clinicians, patients, and other stakeholders, demonstrating that meaningful explanations may vary according to the user and clinical context. This research provides a valuable foundation for developing explainable systems that are technically reliable and practically useful.Collectively, the reviewed literature indicates that explainable AI can contribute to transparency, clinical understanding, and professional trust when implemented appropriately. SHAP, LIME, Grad-CAM, feature importance analysis, and other explanation techniques provide different forms of information about model behavior. However, the available literature also demonstrates that technical explainability does not automatically produce improved patient safety or better clinical outcomes. Differences in explanation quality, user experience, validation methods, and healthcare workflows remain important limitations. The literature therefore supports a human-centered approach in which explainability is evaluated through technical reliability, clinical relevance, professional understanding, and its actual contribution to healthcare decision-making.

METHODOLOGY

This research adopts a qualitative descriptive review methodology to evaluate the role of explainable artificial intelligence systems in enhancing clinical decision-making and patient safety. The study is based on secondary information obtained from peer-reviewed research articles, systematic reviews, meta-analyses, and relevant academic publications. Particular attention is given to literature published between 2025 and 2026 to reflect recent developments in explainable AI, clinical decision support systems, healthcare transparency, and human-centered evaluation. Relevant publications were identified through academic literature searches, including Google Scholar and biomedical and scientific research databases. The selected literature focuses on explanation methods, clinical applications, professional trust, patient safety, model reliability, and barriers to clinical implementation.The reviewed publications were examined according to their relevance to the research objectives and their contribution to understanding explainable AI in healthcare. Particular attention was given to SHAP, LIME, Grad-CAM, attention-based visualization, feature attribution, and other model-interpretation approaches. Studies involving clinical decision support, disease diagnosis, risk prediction, chronic disease management, and healthcare professional evaluation were considered. The review also examined research addressing explanation fidelity, interpretability, usability, fairness, privacy, and clinical validation. The study does not involve original patient recruitment, clinical experiments, or the development and testing of a new AI system. Instead, it provides a structured interpretation of previously published evidence to identify the potential benefits, limitations, and practical implications of explainable AI in healthcare environments.

RESEARCH DESIGN

The research follows a qualitative descriptive review design based on the comparative

examination of existing academic literature. This design is appropriate because the study aims to understand the contribution of explainable AI systems to clinical decision-making and patient safety rather than test a newly developed algorithm. The selected literature is organized according to major themes, including AI transparency, explanation techniques, clinical decision support, healthcare professional trust, patient safety, and implementation challenges. Each theme is examined in relation to the technical and practical requirements of healthcare applications.A comparative approach is also used to examine different explainability methods and their clinical applications. SHAP and LIME are considered in relation to their ability to explain feature contributions in structured clinical data, while Grad-CAM and attention-based visualization are examined for their applications in medical imaging and complex model outputs. The research design further considers differences between technical evaluation and human-centered assessment. Particular attention is given to explanation fidelity, stability, interpretability, usability, and clinical relevance. The design also examines how institutional conditions, professional training, and workflow integration influence the practical adoption of explainable AI systems. This approach provides a comprehensive understanding of explainability as both a technical requirement and a clinical implementation concern.

DATA ANALYSIS

The collected secondary information is analyzed through thematic analysis to identify recurring concepts and findings across the selected publications. The literature is organized into themes related to explanation techniques, clinical transparency, decision-support effectiveness, patient safety, healthcare professional trust, and responsible AI implementation. Each theme is examined to understand how explainable AI methods provide information about the factors influencing automated predictions and recommendations. Particular attention is given to the distinction between explaining model behavior and establishing the clinical correctness of a recommendation.A comparative analysis is used to examine the strengths and limitations of different explainability methods. SHAP is evaluated in relation to feature contribution analysis, while LIME is considered for its ability to approximate individual model predictions through locally interpretable explanations. Grad-CAM and attention-based methods are examined in relation to the visualization of influential image regions and other complex model inputs. The analysis also considers technical evaluation dimensions, including explanation fidelity, stability, consistency, and interpretability. Human-centered evaluation is examined through reported evidence concerning clinician understanding, trust, usability, workflow compatibility, and perceived clinical value. These dimensions are considered separately because a technically faithful explanation may still be difficult to understand or use in clinical practice.The analysis further examines the relationship between explainability and patient safety. Relevant considerations include the identification of influential clinical variables, the recognition of potentially inappropriate AI recommendations, the risk of automation bias, and the importance of professional oversight. Ethical and organizational factors are also examined, including patient

privacy, algorithmic fairness, data quality, accountability, and institutional readiness. The findings are interpreted according to the quality and limitations of the available evidence. This approach ensures that the article distinguishes the potential contribution of explainability from proven improvements in patient outcomes and avoids presenting secondary literature findings as original clinical results.

FINDINGS

The reviewed literature indicates that explainable artificial intelligence can improve the transparency and interpretability of AI-supported clinical decision systems. Techniques such as SHAP, LIME, Grad-CAM, and attention-based visualization provide different ways of examining the information influencing model predictions. These approaches can help healthcare professionals identify important clinical variables, understand image-based predictions, and examine the reasoning associated with individual risk assessments. Their usefulness depends on selecting an appropriate explanation method for the model architecture, data type, and clinical task. The evidence therefore suggests that explainability can support clinical understanding when explanations are technically reliable and relevant to the decision being considered (Abbas et al., 2025; Esfahani et al., 2025).A second important finding concerns the contribution of human-centered explanation design to healthcare professional trust. The reviewed literature indicates that clinicians are more likely to consider AI-supported systems useful when explanations are understandable, clinically relevant, and compatible with existing workflows. Training, usability, professional autonomy, and confidence in system validation also influence acceptance. Complex visualizations or lengthy explanations may increase cognitive workload and reduce practical usefulness, particularly in time-sensitive clinical settings. These findings suggest that explainable AI systems should be designed around the information requirements of their intended users rather than relying exclusively on technical explanation methods (Tun et al., 2025; Salimparsa et al., 2025).

The findings also demonstrate that explainability may support patient safety by helping clinicians examine the factors behind automated recommendations and identify possible inconsistencies. However, explanation quality must be distinguished from predictive accuracy. A system may produce an explanation that appears understandable while still relying on incomplete data, biased patterns, or incorrect predictions. The reviewed evidence identifies explanation fidelity, interpretability, and clinical value as important dimensions of evaluation. It also indicates that prospective clinical assessment and meaningful participation by healthcare professionals are necessary to determine whether explanations improve decision quality in actual healthcare environments (Bauer & Michalowski, 2025; Roychowdhury et al., 2025).Another major finding is that explainable AI continues to face important challenges involving external validation, algorithmic fairness, privacy, accountability, and clinical integration. Existing research demonstrates substantial progress in developing explanation methods, but the evidence

supporting their routine use in healthcare remains limited. Differences in patient populations, medical data quality, clinical specialties, and institutional workflows can affect both model performance and explanation usefulness. The reviewed literature therefore supports the development of standardized evaluation approaches that combine technical reliability with clinical utility and stakeholder requirements. Overall, explainable AI has considerable potential to strengthen trustworthy clinical decision support, but it should complement rather than replace professional judgment and established patient-safety procedures (Zhang et al., 2026; Abbas et al., 2025).

CONCLUSION

Explainable artificial intelligence represents an important development in healthcare because it addresses the transparency and interpretability limitations associated with complex AI-based clinical decision systems. By providing information about the factors influencing automated predictions and recommendations, explainability methods can help healthcare professionals better understand AI outputs and evaluate their relevance to individual patients. Techniques such as SHAP, LIME, Grad-CAM, and attention-based visualization offer different approaches for examining model behavior across structured clinical data, medical imaging, and other healthcare applications. The reviewed literature indicates that these techniques can support clinical understanding and professional trust when they are appropriately selected and evaluated.However, explainability alone cannot guarantee diagnostic accuracy, safe treatment decisions, or improved patient outcomes. The reliability of AI-supported healthcare depends on data quality, model validation, explanation fidelity, fairness, usability, and the ability of healthcare professionals to critically assess automated recommendations. Poorly designed explanations may increase cognitive workload, create misplaced confidence, or encourage inappropriate reliance on AI systems. Future research should therefore prioritize human-centered system design, independent clinical validation, standardized explanation evaluation, professional training, and stronger patient privacy protections. Collaboration among healthcare professionals, AI researchers, institutional administrators, and regulatory stakeholders will be essential for ensuring responsible implementation. In conclusion, explainable AI can contribute to more transparent and trustworthy clinical decision-making, but its successful application must remain grounded in clinical evidence, professional oversight, and patient safety principles.

REFERENCES

Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. Healthcare, 13(17), Article 2154.

Bauer, J. M., & Michalowski, M. (2025). Human-centered explainability evaluation in clinical decision-making: A critical review of the literature. Journal of the American Medical Informatics Association, 32(9), 1477–1484.

Esfahani, H. H., Toyonaga, S., & Oyibo, K. (2025). The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review. Digital Health, 11.

Roychowdhury, S., Lanfranchi, V., & Mazumdar, S. (2025). Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review. Computers in Biology and Medicine, 197(Part A), Article 110944.

Salimparsa, M., Sedig, K., Lizotte, D. J., Abdullah, S. S., Chalabianloo, N., & Muanda, F. T. (2025). Explainable AI for clinical decision support systems: Literature review, key gaps, and research synthesis. Informatics, 12(4), Article 119.

Tun, H. M., Rahman, H. A., Naing, L., & Malik, O. A. (2025). Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review. Journal of Medical Internet Research, 27, Article e69678.

Zhang, K., Wang, D., Lin, F., Xie, J., & Zhou, W. (2026). A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. iScience, 29(3), Article 115026.