Journal of Computer and Communications, 2026, 14(6), 71-86
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Opportunities and Challenges of Explainable AI (XAI) in Health Care: A Review

Mahbuba Begum[1*], Jannatul Ferdush[2]

[1]Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh
[2]Department of Computer Science and Engineering, Jashore University of Science and Technology, Jashore, Bangladesh
Email: *mahbubacse@mbstu.ac.bd, jannatulferdush@just.edu.bd

How to cite this paper: Begum, M. and Ferdush, J. (2026) Opportunities and Challenges of Explainable AI (XAI) in Health Care: A Review. *Journal of Computer and Communications*, **14**, 71-86.
https://doi.org/10.4236/jcc.2026.146006

Received: April 19, 2026
Accepted: June 21, 2026
Published: June 24, 2026

Abstract

The use of Artificial Intelligence (AI) as a decision-making model in healthcare applications faces difficulties because of the black box nature of deep learning (DL) models. Doctors must evaluate a patient's condition based on logical explanations. Therefore, explainable Artificial Intelligence (XAI) plays an important role by providing interpretable explanations for model decisions. This review examines the role of XAI in healthcare by highlighting key techniques, XAI opportunities, application, critical challenges, and recent developments. Also, this research proposes a XAI framework for healthcare system based on existing challenges. Finally, the review outlines practical strategies and future research directions to mitigate these challenges for enhancing trust and usability of XAI systems in real-world healthcare environments.

Keywords

Artificial Intelligence, Trustworthy, Explainability, Healthcare

1. Introduction

AI is a very powerful tool for analyzing medical information in the healthcare sector. It helps medical professionals for predicting and recommending appropriate treatments. For improving healthcare services, smart healthcare involves advanced technologies, such as AI, the Internet of Things (IoT), and cloud computing that uses smart devices. Those devices collect healthcare data like, physical activity, heart rate, blood sugar, and sleep habits including others. Therefore, people can monitor their own health record and share with doctors to diagnose diseases more accurately. AI uses these health records to detect diseases at earlier time, screening health and provides suitable treatment plans.

However, most existing AI models based on black box operations cannot clearly explain their decisions or predictions due to a lack of transparency and potential bias, which makes their use risky when human lives are involved [1] [2]. Researchers argue that AI must be explainable before it can be safely used in high-stakes areas like health care [3]. XAI helps AI models become more transparent and understandable by explaining how and why a particular decision was made. This, in turn, helps doctors, healthcare professionals, and researchers trust AI systems more by increasing transparency, predicting results, and improving current models by identifying errors [1] [2]. When it comes to making decisions about individual patients, current researchers believe that these explanations are mostly false and overly optimistic. They argue that these explainability methods are unreliable and limited which can be incomplete, incorrect, or misleading that often do not help doctors in real understanding [3]. Researchers also found abnormalities in ECG or EEG signals and highlights doctor's notes in clinical tests which require more attention. Recent studies used both Machine Learning (ML) and DL models along with most commonly used XAI methods like: SHapley Additive Explanations (SHAP), Local Interpretable Modelagnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (GradCAM). Overall, SHAP is the most widely used XAI method for identifying important clinical features when predicting diseases or patient outcomes. It is commonly combined with machine learning models such as XGBoost and Random Forest (RF), which work well with structured clinical data. Grad-CAM explains medical image predictions by highlighting important regions in images using heatmaps generated from DL models like Convolutional Neural Networks (CNNs) [4].

This paper reviews the opportunities and challenges of XAI in healthcare, highlighting current methods, applications, limitations, and future research directions. This review includes coverage period (2020-2025), accuracy, and XAI approaches (SHAP, LIME, Grad-CAM etc.) along with applications.

Our research contributions are as follows:

- We present a comprehensive and systematic review of XAI techniques in healthcare applications along with ML and DL models for various clinical data types.
- We highlight opportunities and challenges associated with XAI in healthcare applications.
- We identify key research gaps for existing XAI methods for clinical text data, and provide some suggessions as future research directions.

The paper is organized as follows: Section 2 describes an overview of the XAI system. Section 3 deals with opportunities of XAI in healthcare system. Section 4 covers recent applications of XAI systems. Section 5 discusses existing challenges of XAI in healthcare. Section 6 highlights recent advancements of XAI in healthcare systems. Section 7 presents the proposed XAI framework for healthcare based on identified research gaps. Finally, Section 8 provides a summary of the research and recommendations for future work.

2. Overview of XAI

Healthcare provides real-time patient monitoring, smart environments with strong privacy protection by using advanced technologies including AI, the Internet of Things (IoT), big data, and others. Some studies exist on XAI in healthcare do not focus enough on how data is analyzed and how model explanations are interpreted, which limits real-world use. The study proposes an XAI-based system architecture for analyzing medical images where federated learning ensures patient's privacy and XAI model is applied to validate model performance. Their experiment shows that XAI is effective model for real-world healthcare applications [5]. XAI models help doctors and healthcare professionals to trust AI system so that it can be more understandable and predictable [6] [7]. This explanation ensures that AI-based conclusions are meaningful and useful for medical practice.

Explanations can be global or local where Global explanations help to understand the overall behavior of an AI model, while local explanations explain the reason behind a specific prediction. This two combination provides deeper insights into the model's decisions. Explainability can be classified into two categories: modelagnostic methods, which work with any machine learning algorithm, and model specific methods, which are designed for a particular type of model [8]. The review [9] shows that knowledge graphs improve the comprehensibility of AI systems in healthcare applications by identifying misleading or false information as well as dangerous drug interactions and responses. This knowledge helps bridge the gap between medical experts and AI models. Explanations can be classified as post-hoc or pre-hoc methods, where the post-hoc explanations describe the outcomes generated by an AI system, whereas pre-hoc explanations describe how the AI system functions during the decision-making process [10] [11]. In healthcare, AI systems are evaluated using measures like sensitivity and specificity with higher accuracy requirements than other fields, but in clinical practice, accuracy alone is not sufficient. Doctors need to understand how and why an AI system makes its predictions, as clear explanations help them trust the system and apply its results in real clinical practice. Questions about patient data and features are important for providing clear explanations, although AI models may produce unreliable or unfair predictions for certain training data due to bias [12]. Therefore, the paper [13] identifies that the AI systems should be more explainable and transparent for healthcare professionals.

Another core term interpretability helps developers understand how an AI model makes decisions, while explainability presents these decisions in an understandable way to end-users to build trust, together enabling insight into black-box models and supporting a balance between accuracy and transparency [14]. XAI helps to make an AI application more transparent and also assists in the improvement of the AI application by applying simple cross-domain tools and techniques [15].

For highstake domain like healthcare, transparency is a major issue that allows patients and clinicians to build trust about how and why the system produces a

specific output for given inputs. It ensures reliability and accuracy of the AI system and guarantees human-AI collaboration [16].

Fidelity explain how well an explanation affects the true decision-making process. Low fidelity hides the true effect which is dangerous [17].

Common XAI Techniques and Approaches

This section categorizes commonly used XAI methods. The research [18] reviews popular XAI methods such as LIME, SHAP, and DeepLIFT based on organizations about when explanations are given. This helps users choose the right explanation method for different applications. Based on human thinking and cognition, the framework shows how explanations can be designed in a way that helps people build trust in AI systems.

In intrinsic interpretability, AI models are designed in a transparent way so that people can easily understand how decisions are made without needing extra explanation tools [19]. Some examples of intrinsically interpretable AI models include decision trees, which display decisions step-by-step like a flowchart; generalized additive models (GAMs), which make it evident how each input feature affects the outcome; and rule-based systems, which make decisions based on explicit and predetermined rules. These models are easy to understand because their decision process is visible. This intrinsic model ensure high transparency [20]. Post-hoc explainability methods explain AI models after they have already made a decision, especially complex black-box models. For example: LIME, SHAP, and Layer-wise Relevance Propagation (LRP), where LIME explains a single prediction by slightly changing the input data. Then, it observes how the AI's output changes and creates a simple model that helps people understand why a specific decision was made. Another method, SHAP, explains an AI prediction by assigning a clear importance score to each input feature, showing how strongly each factor influenced the final decision [21]. Deep learning models like LRP explains decisions by tracking the prediction backward through the network layers and highlights the most important features. Although post-hoc methods work well with complex models such as deep neural networks, their explanations can sometimes result in misleading insights [22].

3. Opportunities of XAI in Healthcare

This section highlights how XAI enhances healthcare systems.

3.1. Clinical Decision Support Systems (CDSS)

CDSS are computer-based tools that assist physicians in selecting the best antibiotics for their patients. In addition to helping antimicrobial stewardship programs identify patients who might require treatment modifications, they assist doctors in making safer and more effective antimicrobial treatment decisions [23]. An explainable CDSS also assist doctors in identifying women who are at risk for gestational diabetes mellitus (GDM) [24]. The use of XAI makes the models more

transparent in clinical decision making although some real problems arise with implementation and operational speed [25].

3.2. Trust and Adoption of XAI in Healthcare

XAI applications should be transparent and understandable in making decisions for healthcare applications [26]. For obtaining user trust including XAI, requires transparency [27], specially in healthcare sector. Due to limitations of black-box models, XAI systems increases transparency and lowers the possibility of making poor decisions by giving users insight into the decision-making process [7] [28]-[30]. This trust increases the user acceptance in healthcare applications [31].

3.3. Error Detection and Bias Identification of XAI in Healthcare

XAI helps healthcare stakeholders to understand and trust AI so that decisions are non-biased and accurate [32]. XAI explains how decisions are made and the errors and biases can be found and fixed through reviewing and auditing these explanations [33].

3.4. Regulatory Compliance of XAI in Healthcare

Due to "black-box" nature, AI models faces challenges for regulatory compliance. XAI should be appropriately explainable in building trust on AI systems. These explanation ensures regulatory compliance through emphasizing transparency, accountability, and patient safety for taking clinical decisions. These regulations focus on ensuring the ethical and responsible use of AI [34].

3.5. Knowledge Discovery of XAI in Healthcare

XAI enhances transparency and builds trust among stakeholders in the healthcare sector [35]. Knowledge discovery is a crucial process for identifying patient-related risks and other insights that cannot be detected through traditional analytical approaches. XAI supports the monitoring of personal health and the prediction of healthcare conditions [36].

The related opportunities are illustrated in **Figure 1**.

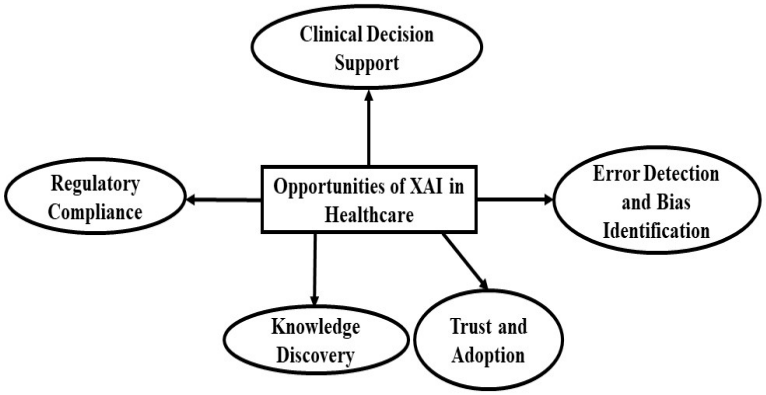

Figure 1. Opportunities of XAI in healthcare.

4. Recent Applications of XAI in Healthcare

Currently, XAI is extensively used in healthcare applications, including medical imaging, assessment of children's developmental status, AI safety, electronic health records, medical text processing, personalized treatment planning, risk prediction, disease diagnosis, COVID-19 management, clinical decision support systems, hospital admission prediction, drug response prediction, and pain recognition, as illustrated in Figure 2 [37]-[42].

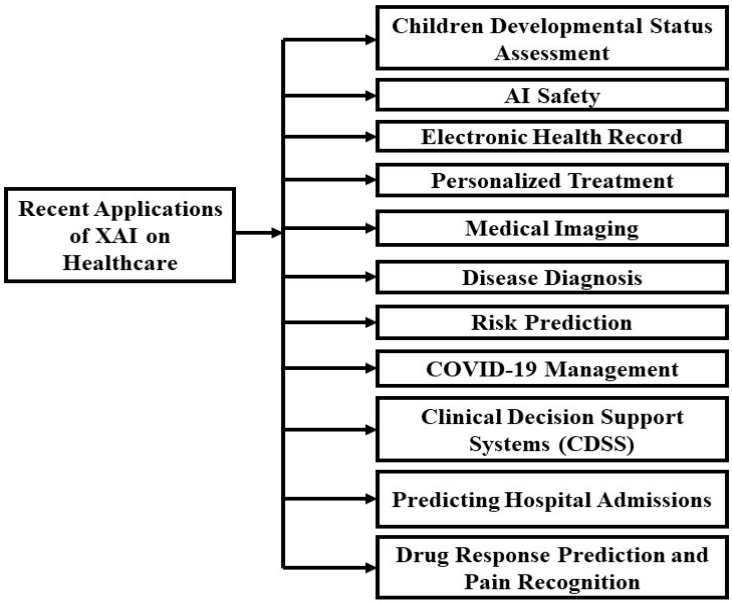

Figure 2. Recent applications of XAI in healthcare.

5. Challenges of XAI in Healthcare

5.1. Trade-Off between Accuracy and Interpretability

It is very difficult to maintain a better trade-off between explainability and predictive accuracy [43]. The explanations should be designed and customized carefully.

5.2. Healthcare Regulations Framework

In healthcare, transparency is essential as decisions of XAI have a direct impact on people's lives [44]. Researchers and developers should analyze XAI compliances with healthcare regulations, ethical standards, and regulatory frameworks [45].

5.3. Explanations Reliability

Various types of explanations can influence physicians' treatment choices. This finding [46] suggests that current ML tools are not sufficiently reliable for recommending improved treatment plans, which may lead to incorrect clinical decisions.

5.4. Lack of Clinical Validation

In healthcare, clinical validation is a primary requirement for the regulatory and certification processes and CDSSs [47]. To achieve clinical validation, prediction accuracy is important. But, AI systems often produce false positive or false negative predictions due to random errors [48].

5.5. Complexity Analysis and Sacalbility Issue

Current XAI models are often very complex, which makes their explanations difficult for users to understand. In addition, generating explanations for large models and datasets requires high computational effort and is usually slow. For instance, Kernel SHAP is a modelagnostic technique that computes explanations by sampling many combinations of input features. This process becomes computationally expensive for high-dimensional data, reducing its scalability for large datasets and complex models [49].

5.6. Human-AI Interaction Issues

The Human-AI Interaction challenge in XAI deals with the explanations produced by AI systems that enable users to interact with and make decisions based on AI outputs.

Figure 3 illustrates the possible challenges of XAI in Healthcare applications.

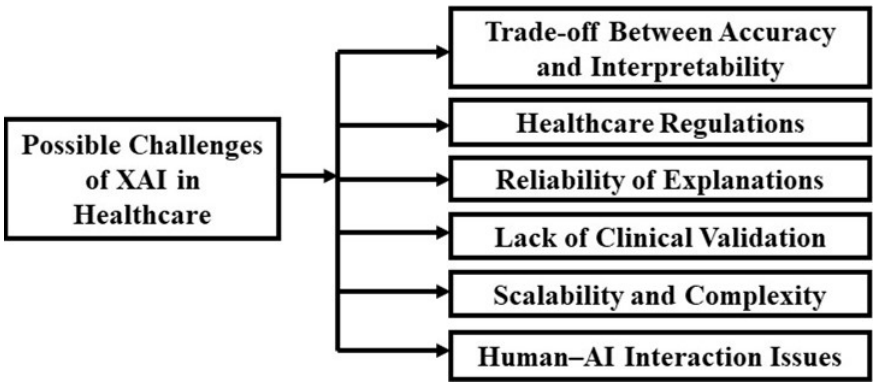

Figure 3. Possible challenges of XAI in healthcare.

In healthcare, explanation quality should be evaluated based on fidelity, robustness, clinician usefulness, and fairness. The system should be robust and secure for avoiding manipulation of predicting explanations that misleads the healthsector. Also, model's biasness and excessive trust on explanations could mislead the system [50]-[55]. These criteria ensure that explanations are accurate, consistent, meaningful for clinicians, and free from bias, thereby supporting safe and effective clinical deployment.

6. Recent Developments of XAI in Healthcare

The work [56] estimates volumetric breast density from images without the need

for segmentation, using a 3D convolutional neural network (CNN). The model achieved strong agreement with minimal bias. SHAP-based explanations showed that accurate predictions relied on relevant breast tissues and provides reliable breast density estimation along with meaningful explainability. XAI addresses several existing challenging issues like model transparency and ethical concerns, by improving the interpretability and trustworthiness of AI models in healthcare. The study [5] proposes an XAI-driven Healthcare 5.0 architecture and validates its effectiveness through case studies on medical imaging and privacy-preserving electrocardiogram (ECG) monitoring using federated learning technique. The method [57] proposes a healthcare framework that integrates AI, blockchain, and the metaverse for designing an efficient digital healthcare services. Doctors and patients interact via a blockchain-based system, where medical data are securely stored and analyzed using XAI techniques such as GradCAM and LIME. The framework ensures reliable disease detection, data protection, transparency, and interpretability in healthcare systems.

The method [58] uses SHAP and LIME to analyze symptomps and predict severity for COVID-19 data. This study shows their effectiveness in supporting explainability. Also, the study [59] uses centralized and federated learning techniques for classifying heart diseases. Centralized models achieved up to 81.1% accuracy using Naive Bayes (NB), while federated logistic regression (LR) reached 78.2% accuracy, protecting patient privacy. Model predictions are interpretated using SHAP and LIME for finding the potentiality of interpretable heart disease prescreening systems. In order to enhance the heart disease detection system, another study [60] presents a hybrid framework that combines explainable AI, deep learning, and machine learning methods. The framework reduces classification error by 20% - 25% across multiple datasets. It addresses data imbalance, missing values, and feature inconsistencies. This work enhances clinical trust and scalability in healthcare systems. Also, this work delivers high predictive performance for the healthcare systems.

The work [61] integrates Decision Trees (DT), NB, RF, and XGBoost for improving both accuracy and interpretability in predicting the risks of diseases including Diabetes, Anaemia, Thalassemia, Heart Disease, and Thrombocytopenia. The method achieves 99.2% accuracy while model predictions are achieved through SHAP and LIME. The study [37] introduces PersonalCareNet, a new deep learning framework for addressing the lack of interpretability in existing AI healthcare models. It delivers both global and patient-specific explanations by combining CNN with attention mechanisms and SHAP. Also, the study [62] predicts workplace mental health using XAI and ML methods including RF, xGBoost, SVM, and AdaBoost. The xGBoost and RF performed best and achieved high accuracy, while SHAP and LIME provided transparent explanations of some important factors. It includes treatment-seeking behavior and past or present mental health conditions.

This review divides studies into some significant categories such as:

- Used Models: CNN, Random Forest, and XGBoost.
- Used XAI: SHAP and LIME.

These are used as current research methods that find the most important clinical features for enhancing diagnostic transparency.

- Clinical tasks: Disease diagnosis, risk prediction, drug response prediction, and mental health analysis.
- Data types: Medical imaging, structured clinical data, ECG/EEG signals, and text data.
- Also, existing methods find the risk affecting factors for the patient.

Table 1. Summary of the existing XAI methods in healthcare.

References, year	Used Methods	XAI Approaches	Accuracy	Healthcare Applications
H. M. van der Velden *et al.* [56], 2020	CNN	SHAP	-	Breast density estimation
D. Saraswat *et al.* [5], 2022	CNN, Federated transfer learning	CAM and Grad-CAM	98%	COVID-19 patients
S. Ali et al. [57], 2023	Blockchain	Grad-CAM, LIME	-	Healthcare
A. Nambiar *et al.* [58], 2023	DT, XGBoost Classifier, and Neural Network Classifier	SHAP, LIME	-	COVID-19 symptom analysis and severity prediction
Rodriguez and Nafea [59], 2025	Linear-kernel SVM model	SHAP	83.3%	Cardiovascular disease detection
Talukder *et al.* [60], 2025	Multilayer Perceptron (MLP)	SHAP, LIME	100.0%	Heart disease detection
Agrawal *et al.* [61], 2025	DT, RF, NB	XGBoost, SHAP, LIME	99%	Healthcare
M. S. Vani *et al.* [36], 2025	CNN + Attention	SHAP, Grad-CAM, Force Plot, Feature Importance	97.86%	Healthcare
T. Mokheleli *et al.* [62], 2025	RF, XGBoost, SVM, and AdaBoost	SHAP, LIME	94%	Mental health

Most of the existing XAI approaches likely SHAP, LIME, and GradCAM offer post-hoc explanations. These explanations might not fully reflect the true internal decision-making processes of complex models. Misinterpretation can have serious consequences in healthcare sector which raises the concerns about explanation fidelity. As a result, we propose an XAI framework for healthcare in Section 7 that aims to minimise these existing challenges.

We have discussed existing challenges at Section 5. Based on Table 1, we have summarized some limitations:

- Strong clinical validation,
- Cross validation,
- Bias and overfitting,
- And, also, rely on post-hoc explanations.

7. Proposed Framework for XAI in Healthcare

Figure 4 presents the proposed framework for an explainable AI-based healthcare system.

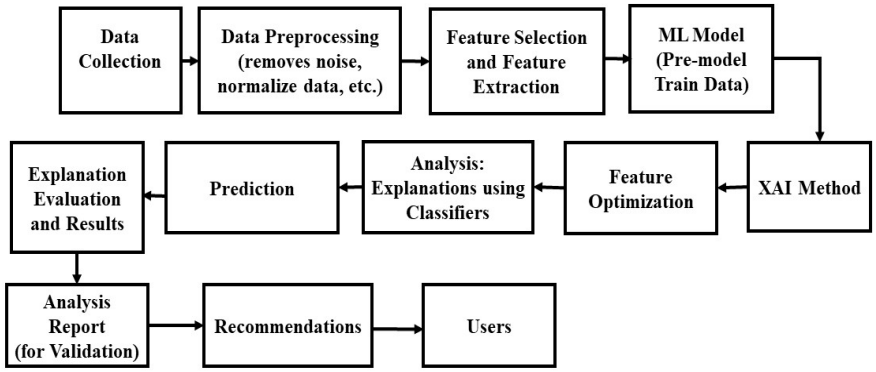

Figure 4. Proposed framework for XAI in healthcare.

This framework addresses the major challenges discussed in Section 5, including the trade-off between accuracy and interpretability, lack of clinical validation, reliability of explanations, scalability issues, regulatory concerns, and human-AI interaction limitations.

The framework starts by collecting the raw healthcare data from clinical databases, medical sensors, electronic health records, and medical imaging devices. Healthcare data are frequently noisy, incomplete, and heterogeneous. Therefore, the preprocessing stage reduces bias and enhances data quality. For this, it performs several operations likely noise reduction, normalization, missing value handling, and data cleaning. This stage directly addresses the challenges of unreliable predictions and model biasness discussed in Section 5.

After preprocessing, feature selection and feature extraction are applied to identify clinically significant attributes. It also reduces dimensionality and computational complexity. This stage improves scalability and reduces the complexity of explanation generation, which helps overcome the scalability and complexity challenges of existing XAI models. Furthermore, selecting clinically meaningful features improves explanation interpretability for healthcare professionals.

The processed features are then used to train machine learning or deep learning models using historical and labeled clinical data. To reduce overfitting and improve generalization performance, cross-validation and performance evaluation are incorporated during model training. This stage addresses the challenge of balancing predictive accuracy and interpretability while improving model robustness and reliability. The optimized features analysis the explanations using classifiers and those explanations guide the generation of model predictions. The explanation evaluation and result validation ensure reliability, fairness, and clinical relevance.

This step helps reduce false-positive and false-negative predictions. It also addresses the lack of clinical validation discussed in Section 5. The validated results

generate report that summarizes predictions, explanations, and performance metrics in an interpretable format. Also, this report provides practical advices which supports clinicians and decision-makers in diagnosis, patient management, and treatment planning. The completed results are then delivered to end users.

These steps address the human-AI interaction challenges. Also, it guarantees that explanations remain relevant and trustworthy in real clinical settings. The proposed framework ensures that predictions are not only accurate but also transparent, reliable, and helpful for medical applications by integrating explainability throughout the ML lifecycle.

8. Conclusions and Future Directions

XAI plays a significant role in the healthcare industry. It enhances accountability, transparency, and trust in AI-driven systems. It provides interpretable predictions and clear explanations that enable clinicians and healthcare professionals for validating and integrating the AI system into clinical decision-making. XAI supports informed diagnoses, treatment planning, and patient management. But, there exist several challenges on this area including the trade-offs between interpretability and predictive accuracy, limited clinical relevance of some explanations, data quality and bias issues, and difficulties in integrating XAI tools into existing clinical methods. Also, regulatory, ethical, and validation-related issues further complicate real-world deployment. Future work should focus on hybrid modeling techniques that maintain a better trade-off between explainability and performance, clinician-centered system design, reliable explanations, bias minimization, standard framework, proper clinical validation, cloud-based healthcare, and authorized healthcare regulations for ensuring trustworthy and responsible healthcare system.

Author Contributions

All authors have contributed equally to performing this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dave, D., Naik, H., Singhal, S. and Patel, P. (2020) Explainable AI Meets Healthcare: A Study on Heart Disease Dataset. arXiv:2011.03195.

[2] Pawar, U., O'Shea, D., Rea, S. and O'Reilly, R. (2020) Explainable AI in Healthcare. 2020 *International Conference on Cyber Situational Awareness, Data Analytics and Assessment* (*CyberSA*), Dublin, 15-19 June 2020, 1-2. https://doi.org/10.1109/cybersa49311.2020.9139655

[3] Ghassemi, M., Oakden-Rayner, L. and Beam, A.L. (2021) The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, **3**, e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

[4] Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F. and 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

[5] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V.K., Tanwar, S., Sharma, G., *et al.* (2022) Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*, **10**, 84486-84517. https://doi.org/10.1109/access.2022.3197671

[6] Wohlin, C. (2014) Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. *Proceedings of the* 18*th International Conference on Evaluation and Assessment in Software Engineering*, London, 13-14 May 2014, 1-10. https://doi.org/10.1145/2601248.2601268

[7] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., *et al.* (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, **58**, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[8] Adadi, A. and Berrada, M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6**, 52138-52160. https://doi.org/10.1109/access.2018.2870052

[9] Rajabi, E. and Kafaie, S. (2022) Knowledge Graphs and Explainable AI in Healthcare. *Information*, **13**, Article 459. https://doi.org/10.3390/info13100459

[10] Kenny, E.M., Ford, C., Quinn, M. and Keane, M.T. (2021) Explaining Black-Box Classifiers Using Post-Hoc Explanations-By-Example: The Effect of Explanations and Error-Rates in XAI User Studies. *Artificial Intelligence*, **294**, Article 103459. https://doi.org/10.1016/j.artint.2021.103459

[11] Botana, I.L., Eiras-Franco, C. and Alonso-Betanzos, A. (2020) Regression Tree Based Explanation for Anomaly Detection Algorithm. 3*rd XoveTIC Conference Proceedings* 2020, **54**, 7. https://doi.org/10.3390/proceedings2020054007

[12] Yang, C.C. (2022) Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*, **6**, 228-239. https://doi.org/10.1007/s41666-022-00114-1

[13] Elul, Y., Rosenberg, A.A., Schuster, A., Bronstein, A.M. and Yaniv, Y. (2021) Meeting the Unmet Needs of Clinicians from AI Systems Showcased for Cardiology with Deep-Learning-Based ECG Analysis. *Proceedings of the National Academy of Sciences*, **118**, e2020620118. https://doi.org/10.1073/pnas.2020620118

[14] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G. (2019) XAI—Explainable Artificial Intelligence. *Science Robotics*, **4**, eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[15] Wang, Y.-C., Chen, T.-C.T. and Chiu, M.-C. (2023) An Improved Explainable Artificial Intelligence Tool in Healthcare for Hospital Recommendation. *Healthcare Analytics*, **3**, Article 100147. https://doi.org/10.1016/j.health.2023.100147

[16] Gunning, D. (2017) Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA).

[17] Sokol, K. and Flach, P. (2020) Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 56-67. https://doi.org/10.1145/3351095.3372870

[18] Arunraju Chinnaraju, (2025) Explainable AI (XAI) for Trustworthy and Transparent Decision-Making: A Theoretical Framework for AI Interpretability. *World Journal*

of Advanced Engineering Technology and Sciences, **14**, 170-207. https://doi.org/10.30574/wjaets.2025.14.3.0106

[19] Vale, D., El-Sharif, A. and Ali, M. (2022) Explainable Artificial Intelligence (XAI) Post-Hoc Explainability Methods: Risks and Limitations in Non-Discrimination Law. *AI and Ethics*, **2**, 815-826. https://doi.org/10.1007/s43681-022-00142-y

[20] Mota, B., Faria, P., Corchado, J. and Ramos, C. (2024) Explainable Artificial Intelligence Applied to Predictive Maintenance: Comparison of Post-Hoc Explainability Techniques. In: Longo, L., Lapuschkin, S. and Seifert, C. Eds., *Communications in Computer and Information Science*, Springer, 353-364. https://doi.org/10.1007/978-3-031-63803-9_19

[21] Narkhede, J. (2024) Comparative Evaluation of Post-Hoc Explainability Methods in AI: LIME, SHAP, and Grad-Cam. 2024 4*th International Conference on Sustainable Expert Systems* (*ICSES*), Kaski, 15-17 October 2024, 826-830. https://doi.org/10.1109/icses63445.2024.10762963

[22] Belaid, M.K., Bornemann, R., Rabus, M., Krestel, R. and Hüllermeier, E. (2023) Compare-xAI: Toward Unifying Functional Testing Methods for Post-Hoc XAI Algorithms into a Multi-Dimensional Benchmark. In: Longo, L., Ed., *Communications in Computer and Information Science*, Springer, 88-109. https://doi.org/10.1007/978-3-031-44067-0_5

[23] Simon, M.S. and Calfee, D.P. (2017) Optimizing the Use of Antimicrobial Agents: Anti-Microbial Stewardship and Outpatient Parenteral Antimicrobial Therapy (OPAT). In: Cohen, J., Powderly, W.G. and Opal, S.M., Eds., *Infectious Diseases*, Elsevier, 1197-1202.E1. https://doi.org/10.1016/b978-0-7020-6285-8.00139-8

[24] Du, Y., Rafferty, A.R., McAuliffe, F.M., Wei, L. and Mooney, C. (2022) An Explainable Machine Learning-Based Clinical Decision Support System for Prediction of Gestational Diabetes Mellitus. *Scientific Reports*, **12**, Article No. 1170. https://doi.org/10.1038/s41598-022-05112-2

[25] Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N. and Kroeker, K.I. (2020) An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. *npj Digital Medicine*, **3**, Article No. 17. https://doi.org/10.1038/s41746-020-0221-y

[26] Sadeghi, Z., Alizadehsani, R., CIFCI, M.A., Kausar, S., Rehman, R., Mahanta, P., *et al.* (2024) A Review of Explainable Artificial Intelligence in Healthcare. *Computers and Electrical Engineering*, **118**, Article 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

[27] Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G. and De Hert, P. (2022) Bridging the Gap between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine*, **17**, 72-85. https://doi.org/10.1109/mci.2021.3129960

[28] Miller, T. (2019) Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, **267**, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[29] Kostopoulos, G., Davrazos, G. and Kotsiantis, S. (2024) Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics*, **13**, Article 2842. https://doi.org/10.3390/electronics13142842

[30] Liao, Q.V., Gruen, D. and Miller, S. (2020) Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the* 2020 *CHI Conference on Human Factors in Computing Systems*, Honolulu, 25-30 April 2020, 1-15. https://doi.org/10.1145/3313831.3376590

[31] Kovari, A. (2024) AI for Decision Support: Balancing Accuracy, Transparency, and Trust across Sectors. *Information*, **15**, Article 725. https://doi.org/10.3390/info15110725

[32] Allgaier, J., Mulansky, L., Draelos, R.L. and 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**, Article 102616. https://doi.org/10.1016/j.artmed.2023.102616

[33] Lipton, Z.C. (2016) The Mythos of Model Nterpretability. *Communications of the ACM*, **61**, 36-43.

[34] Gupta, N. (2025) Explainable AI for Regulatory Compliance in Financial and Healthcare Sectors: A Comprehensive Review. *International Journal of Advances in Engineering and Management*, **7**, 489-494. https://doi.org/10.35629/5252-0703489494

[35] Nahavandi, S., Alizadehsani, R., Nahavandi, D., Mohamed, S., Mohajer, N., Ro-Konuzzaman, M. and Hossain, I. (2022) A Comprehensive Review on Autonomous Navigation. arXiv:2212.12808.

[36] Vani, M.S., Sudhakar, R.V., Mahendar, A., Ledalla, S., Radha, M. and Sunitha, M. (2025) Personalized Health Monitoring Using Explainable AI: Bridging Trust in Predictive Healthcare. *Scientific Reports*, **15**, Article No. 31892. https://doi.org/10.1038/s41598-025-15867-z

[37] George, R., Ellis, B., West, A., Graff, A., Weaver, S., Abramowski, M., *et al.* (2023) Ensuring Fair, Safe, and Interpretable Artificial Intelligence-Based Prediction Tools in a Real-World Oncological Setting. *Communications Medicine*, **3**, Article No. 88. https://doi.org/10.1038/s43856-023-00317-6

[38] Zhang, H. and Ogasawara, K. (2023) Grad-Cam-Based Explainable Artificial Intelligence Related to Medical Text Processing. *Bioengineering*, **10**, Article 1070. https://doi.org/10.3390/bioengineering10091070

[39] Gouverneur, P., Li, F., Shirahama, K., Luebke, L., Adamczyk, W.M., Szikszay, T.M., *et al.* (2023) Explainable Artificial Intelligence (XAI) in Pain Research: Understanding the Role of Electrodermal Activity for Automated Pain Recognition. *Sensors*, **23**, Article 1959. https://doi.org/10.3390/s23041959

[40] Sada Del Real, K. and Rubio, A. (2023) Discovering the Mechanism of Action of Drugs with a Sparse Explainable Network. *eBioMedicine*, **95**, Article 104767. https://doi.org/10.1016/j.ebiom.2023.104767

[41] Park, A., Lee, Y. and Nam, S. (2023) A Performance Evaluation of Drug Response Prediction Models for Individual Drugs. *Scientific Reports*, **13**, Article No. 11911. https://doi.org/10.1038/s41598-023-39179-2

[42] Drobnič, F., Starc, G., Jurak, G., Kos, A. and Pustišek, M. (2023) Explained Learning and Hyperparameter Optimization of Ensemble Estimator on the Bio-Psycho-Social Features of Children and Adolescents. *Electronics*, **12**, Article 4097. https://doi.org/10.3390/electronics12194097

[43] Kose, U., Sengoz, N., Chen, X. and Saucedo, J.A.M. (2024) Explainable Artificial Intelligence (XAI) in Healthcare. CRC Press.

[44] Kiseleva, A., Kotzinos, D. and De Hert, P. (2022) Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. *Frontiers in Artificial Intelligence*, **5**, Article ID: 879603. https://doi.org/10.3389/frai.2022.879603

[45] Johannssen, A. and Chukhrova, N. (2025) The Crucial Role of Explainable Artificial

Intelligence (XAI) in Improving Health Care Management. *Health Care Management Science*, **28**, 565-570. https://doi.org/10.1007/s10729-025-09720-y

[46] Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F. and Gajos, K.Z. (2021) How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection. *Translational Psychiatry*, **11**, Article No. 108. https://doi.org/10.1038/s41398-021-01224-x

[47] Higgins, D. and Madai, V.I. (2020) From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Advanced Intelligent Systems*, **2**, Article 2000052. https://doi.org/10.1002/aisy.202000052

[48] Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020) Fooling LIME and Shap. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, 7-9 February 2020, 180-186. https://doi.org/10.1145/3375627.3375830

[49] Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X. and Wang, T. (2018) Interpretable Deep Learning under Fire. *Proceedings of the* 29*th USENIX Conference on Security Symposiu*, Berkeley, 12-14 August 2020, 1659-1676.

[50] Thames, C. and Sun, Y. (2024) A Survey of Artificial Intelligence Approaches to Safety and Mission-Critical Systems. 2024 *Integrated Communications, Navigation and Surveillance Conference* (*ICNS*), Herndon, 23-25 April 2024, 1-12. https://doi.org/10.1109/icns60906.2024.10550712

[51] Wei, J., Turbé, H. and Mengaldo, G. (2024) Revisiting the Robustness of Posthoc interpretability Methods. arXiv:2407.19683.

[52] Buçinca, Z., Malaya, M.B. and Gajos, K.Z. (2021) To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, **5**, 1-21. https://doi.org/10.1145/3449287

[53] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., *et al.* (2021) Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the* 2021 *CHI Conference on Human Factors in Computing Systems*, Yokohama, 8-13 May 2021, 1-16. https://doi.org/10.1145/3411764.3445717

[54] Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making*, **20**, Article No. 310. https://doi.org/10.1186/s12911-020-01332-6

[55] Aysel, H.I., Cai, X. and Prugel-Bennett, A. (2025) Explainable Artificial Intelligence: Advancements and Limitations. *Applied Sciences*, **15**, Article 7261. https://doi.org/10.3390/app15137261

[56] van der Velden, B.H.M., Janse, M.H.A., Ragusi, M.A.A., Loo, C.E. and Gilhuijs, K.G.A. (2020) Volumetric Breast Density Estimation on MRI Using Explainable Deep Learning Regression. *Scientific Reports*, **10**, Article No. 18095. https://doi.org/10.1038/s41598-020-75167-6

[57] Ali, S., Abdullah, Armand, T.P.T., Athar, A., Hussain, A., Ali, M., *et al.* (2023) Metaverse in Healthcare Integrated with Explainable AI and Blockchain: Enabling Immersiveness, Ensuring Trust, and Providing Patient Data Security. *Sensors*, **23**, Article 565. https://doi.org/10.3390/s23020565

[58] Nambiar, A., S, H. and S, S. (2023) Model-Agnostic Explainable Artificial Intelligence Tools for Severity Prediction and Symptom Analysis on Indian COVID-19 Data. *Frontiers in Artificial Intelligence*, **6**, Article ID: 1272506. https://doi.org/10.3389/frai.2023.1272506

[59] Rodriguez, M.P., Oladipo, E. and Nafea, M. (2025) Centralized and Federated Heart Disease Classification Using UCI Dataset: A Benchmark with Interpretability Analysis. 2025 *IEEE Evolution-Life Members Conference*, Boston, 11-13 June 2025, 1-8. https://doi.org/10.1109/evolution65010.2025.11044926

[60] Talukder, M.A., Talaat, A.S., Kazi, M. and Khraisat, A. (2025) XAI-HD: An Explainable Artificial Intelligence Framework for Heart Disease Detection. *Artificial Intelligence Review*, **58**, Article No. 385. https://doi.org/10.1007/s10462-025-11385-6

[61] Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P. and Hamdare, S. (2025) Fostering Trust and Interpretability: Integrating Explainable AI (XAI) with Machine Learning for Enhanced Disease Prediction and Decision Transparency. *Diagnostic Pathology*, **20**, Article No. 105. https://doi.org/10.1186/s13000-025-01686-3

[62] Mokheleli, T., Bokaba, T. and Mbunge, E. (2025) Explainable Artificial Intelligence for Workplace Mental Health Prediction. *Informatics*, **12**, Article 130. https://doi.org/10.3390/informatics12040130