International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 2, Issue 3 July-September 2025 ISSN: 3048-975X
https://ijrasht.com/

Explainable AI in Healthcare: Models, Applications, and Challenges

Mukesh Kumar, Professor & Research Supervisor, Department of Computer Science and Engineering, NIILM University, Kaithal, Haryana mrana91@gmail.com

Ashish Kumar, Research Scholar, Department of Computer Science and Engineering, NIILM University, Kaithal, Haryana

Abstract: The healthcare industry is changing due to the application of Artificial Intelligence (AI) because it opens the possibility of making predictions in analytics, clinical decision support, medical imaging diagnostics, and custom treatment. Nevertheless, most of the recent AI models, particularly the deep learning models, are classified as black boxes since they are not interpretable. This veil puts the ethics, law and practice of health care in doubt where judgment directly affects patient safety and trust. The concept of explainable artificial intelligence (XAI) has emerged as a significant research field to make AI systems easier to understand and more open, interpretable, and credible by providing explanations of their behaviour at model level. This article is a review of explainable AI in healthcare that covers the types of models, their uses, and challenges. It further describes common techniques, including the post-hoc interpretability techniques (e.g. SHAP, LIME), the inherently interpretable models (e.g. decision trees, rule-based systems), and hybrids. It has also been discussed within the framework of diagnostic imaging, electronic health records (EHRs), drug discovery, and precision medicine. Also, there are still concerns regarding how to deal with trade-offs between accuracy and interpretability, how to measure evaluation metrics in a standardized manner, how to evaluate fairly, and how to translate XAI into clinical practice. The findings of the existing literature indicate that XAI will increase compliance with clinicians and regulators and patient empowerment. But under varying clinical conditions, scalability, the expense and variability of interpretability, continue to be key bottlenecks. The new directions are building the context-specific XAI models, the federated learning support, the alignment of the models with the ethical and legal principles, and GDPR. In this paper, the explainable AI is identified as a key to the responsible use of AI in healthcare. XAI also ensures that AI is used to make healthcare delivery safer, more ethical and more effective by closing the gap between complex models and the way human beings make decisions.

Keywords: Explainable AI, healthcare, interpretability, medical decision support, trustworthy AI.

1. Introduction

Nowadays AI is an unavoidable part of the healthcare system, as it is possible to predict diseases or formulate a personal treatment, allocate resources most efficiently. However, most developed AI systems and in particular deep neural nets are black boxes, meaning they do not give explanations as to why outputs are distributed in a certain way (Doshi-Velez and Kim, 2017). Such interpretability can be a significant detriment to trust, accountability, and adoption by clinicians within the healthcare environment. To be more specific, the diagnostic model, which prescribes a form of treatment and keeps the factors, which necessitate the need to receive this treatment a secret, makes the question of safety and justice rather dubious (Adabi and Berrada, 2018).

We can start with Explainable AI (XAI) as one of the technologies that can help address this challenge and make AI work responsible and understandable to the professional and patient community. XAI is more accountable, compliant with the regulations, and acceptable by the clinicians. Due to the stakes of healthcare being this high, it can be helpful to understand how AI recommendations operate. This paper addresses the model, uses, and issues of XAI within healthcare and how it contributes to the use of sound clinical judgments.

2. Background of the Study

Healthcare is a high-dimensional system with imaging, genomics and electronic health records (EHRs) among others. The second impact of deep learning is that deep learning has improved the quality of prediction and decreased interpretability (Ribeiro et al., 2016). Correct predictions are typically sought by clinicians, as well as rationalizations to defend judgments. In the same way, AI systems that identify tumours should be able to identify single areas of radiographic images that affect predictions (Arrieta et al., 2020).

Figure 1. Levels of Explainability in AI Models for Healthcare

The XAI systems are described as being explicative on more than one level:

- Model-level explainability (interpretable models such as decision trees).

- Post-hoc interpretability (including SHAP, LIME, Grad-CAM).
- Hybrid techniques (trading off between interpretability and accuracy).

Everything needs to be done to reduce the degree of clinician mistrust and adapt AI systems to healthcare requirements and policies (Holzinger et al., 2019).

3. Justification

Both ethical, legal and practical needs of the clinical decision making in the healthcare industry justify explainable AI use in healthcare. There are few areas of life in which the influence of opaque AI system can be so dramatic as in healthcare (Rajpurkar et al., 2017). Other than that, we have other laws and regulations like GDPR which are linked to the right to explain and this is why, we need to be more transparent in the context of automated decisions. This seems more feasible clinically, as AI with explanable reasoning has a higher chance of being trusted and accepted by the clinician (Tonekaboni et al., 2019). The patients also require evident motives to accept AI-related interventions. Therefore, XAI is an essential part of compliance, as well as trust building and patient outcomes.

4. Objectives of the Study

1. To test the usefulness of explainable AI models in healthcare.
2. To investigate XAI uses in diagnostics, treatment planning and medical data analysis.
3. To identify problems in the use of XAI systems in clinical practice.
4. To evaluate the new research opportunities in plausible AI in healthcare.

5. Literature Review

Interpretable Models: decision tree, logistic regression, rule-based, and category are all interpretable models that do not need to be predictive (Caruana et al., 2015).

Post-hoc Methods How to view the black-box models (LIME and SHAP) is usually through the lens of Liberal Learning (Ribeiro et al., 2016). Deep Learning Interpretability: Grad-Cam enables the diagnostic data to inform a health image as salient points will be identified in the image (Selvaraju et al., 2017). Medical care: Rajpurkar et al. (2017) reported the use of deep learning in the diagnosis of chest X-rays, and Holzinger et al. (2019) investigated the interpretability required to make deep learning clinician-acceptable. In the article, Arrieta et al. (2020) even mentioned that XAI is a middle ground between ethics and AI-based models of healthcare delivery.

6. Material and Methodology

The current paper adopts a Systematic Literature Review (SLR) as a method to introduce a systematic and transparent literature review on Explainable Artificial Intelligence (XAI) in healthcare. The goal is to combine the findings of other studies and determine whether interpretability can increase clinical decision support, regulatory compliance and trust in AI-based healthcare solutions.

Sources of Data

Broad search of a database was carried out in order to include both theoretical and applied research in the medical area and in computer science. The following were considered as digital libraries and repositories:

- IEEE Xplore: A technical and algorithmic development of AI and XAI models.
- PubMed: To locate medical and clinical research, particularly that applies XAI to real healthcare data.
- SpringerLink: To find peer-reviewed journal articles and book chapters about the application of AI in the medical sphere.
- ACM Digital Library: To find cross-disciplinary sources of interest in the research area of AI algorithms, user interaction, and explainability.

To mediate AI and clinical practice, to apply AI to medical and healthcare environments, To address applied research in medical and healthcare disciplines, to mediate AI and medical practice.

- Controlled vocabulary as well as Boolean operators were used to locate relevant literature. The following are the keywords and combinations that were used:
- AND: health care + explainable AI
- Including: Medical Decision Support AND Interpretability.
- XAI and Clinical Decision-Making
- Transparent AI + Diagnostics

This ensured the retrieval of studies which addressed both technical design of explainable algorithms and their use in the health care set up.

Timeframe

The search contains only recent articles (2015-2023) because the research topic of explainability has become increasingly popular in the field during the same period, particularly due to the introduction of GDPR and FDA regulations that have made both transparency and responsibility fundamental principles of research in the field.

Selection Criteria

The following inclusion and elimination criteria were used to arrive at quality and relevancy:

Inclusion Criteria

- The papers that suggest, design or test XAI approaches in healthcare.
- Medical (e.g. imaging, EHR, genomics) empirical studies.
- Studies that have focused on interpretability of clinical decision support.

Exclusion Criteria

- Articles, which do not concern healthcare applications in particular.
- Hypothetical fragments of work lacking medical evidence.
- Articles that belonged to the undefined time.

Following procedure was used to make the analysis and extract data:

To conduct a systematic analysis of eligible studies, the following dimensions were applied:

1. XAI Technique: feature attribution, saliency maps, rule-based models, counterfactual explanations etc.
2. Healthcare case study: diagnostics imaging, electronic health records (EHR), predictive analytics, drug discovery.
3. Evaluation Tests: accuracy, Clinician interpretability, Medical compliance.
4. Conclusions: Practical implementation benefits, limitations and failures.

7. Results and Discussion

Synthesis of reviewed literature suggests that there are certain fundamental benefits and concerns of XAI in medicine.

Clinician Trust

One of the most important advantages of XAI is the ability to make machine learning models more understandable to clinicians. Unlike black-box models (e.g. deep neural networks), XAI does not generate outputs that are hard to interpret, such as heatmaps of medical images, decision rules, or feature rankings. Saliency-based medical imaging methods provide an example where radiologists can visualize the margins or the position of a tumor or an abnormal tissue, which is consistent with their intuitive interpretation of the diagnosis. This increases acceptance and use of AI tools by clinical working.

Table 1. Evaluation Dimensions for Reviewed Studies on XAI in Healthcare

Dimension	Description	Examples in Reviewed Studies
XAI Technique	Type of explainability method applied	SHAP, LIME, Decision Trees, Rule-based, Grad-CAM
Healthcare Domain	Clinical context where XAI is applied	Diagnostic Imaging, EHR, Genomics, Drug Discovery
Evaluation Criteria	Metrics used to assess performance and interpretability	Accuracy, Clinician Trust, Interpretability, Compliance
Practical Outcomes	Real-world relevance, benefits, and limitations	Improved diagnostic confidence, regulatory approval, challenges in scalability

Ethics and Standards of Regulatory Compliance

Responsible and transparent healthcare technologies are required by the FDA and European Commission (GDPR, AI Act) and other regulatory organizations. XAI is explicitly supportive of these requirements since it enables the AI systems to justify their predictions and decisions. Studies have shown that the addition of interpretability to clinical decision support systems enhances their acceptance by regulatory authorities and their compliance with ethical standards such as fairness, non-maleficence and accountability.

Better Diagnostic support

XAI has already been applied to supplement diagnostic processes displaying the reasoning behind the AI forecast. Oncology Oncologists are exploring the use of XAI techniques to detect tumors in radiographic images to support their treatment plans. The cardiology field applies rule-based explainable models to find significant biomarkers

or ECG signal patterns that predict heart disease. In genomics, feature attribution algorithms have been demonstrated to model the genes with the best association to disease risk. In this way, using such applications, clinicians will not only be able to accept AI outputs but also be able to cross-verify and challenge predictions to reduce the risks of misdiagnosis.

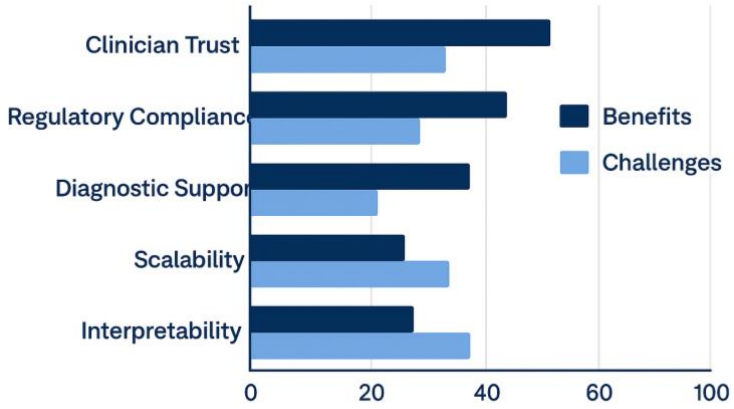

Figure 2: Benefits and Challenges of Explainable AI in Healthcare

Challenges Identified

Despite this progress, there are certain problems:

One is the trade-off between Interpretability and Accuracy: Models that are easier to explain can tend to be more accurate, but black-box more accurate models are also more often used. Technicality must be reduced to the level of clinical user; explanation must be reduced to usability. XAI continues to be linked to large resource requirements when it is carried out on big healthcare data, such as the EHR systems of a variety of hospitals.

Overall Insights

The results indicate that XAI is both a technical improvement and a requirement to trustful AI in healthcare. XAI makes AI-based decision support systems more responsible, more acceptable, and, of course, safer by bridging the gap between machine learning predictions and clinician reasoning.

Challenges include:

- Accuracy vs. interpretability.
- Lack of standard evaluation measures.
- Many healthcare settings.
- Possible risk of false or simplistic explanations.

Thus, more people are likely to adopt and believe in XAI, yet the use of this concept in healthcare remains under-developed.

8. Limitations of the Study

The study problem is that XAI methods are constantly evolving, and, in the near future, the findings may become obsolete (Holzinger et al., 2019). The majority of the studies also work on small-scale, or domain-related data, which are less generalizable. Cross-model comparison is also restricted by a few standard benchmarks.

9. Future Scope

There should be future research regarding:

- Federated XAI: Learning explainability that is privacy-preserving to safely deploy AI to the clinic (Tonekaboni et al., 2019).
- Multimodal XAI: Imaging, genomics and clinical notes combined to view the entire picture.
- Explainability standards: Work out the standards and protocols to evaluate explanations.
- Human value and patient rights as the framework will be offered to the XAI Ethical integration (Arrieta et al., 2020).

10. Conclusion

Explainable AI is needed to enable responsible healthcare innovation. It is useful in removing the gap between what performance models are capable of and what humans are capable of choosing through transparency, trust

and accountability. XAI will result in safe implementation of AI in clinical practice, despite technical and operational setbacks, which complies with ethical requirements and regulatory requirements.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.
4. Chen, J., Song, L., & Wainwright, M. J. (2018). Learning to explain: An information-theoretic perspective on model interpretability. *ICML 2018*.
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
6. Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*.
7. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2019). What do we need to build explainable AI systems for the medical domain? *Review in Methods of Information in Medicine*, 58(4–5), e1–e6.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
9. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
11. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITW 2017 IEEE Information Theory Workshop*, 1–10.
12. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *ICCV 2017*.
13. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable AI for clinical end use. *Machine Learning for Healthcare Conference*, 359–380.
14. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
15. Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*.