IJIS: Vol.1, Issue 11, December 2025, Page: 24-29 ISSN: 3049-3277

International Journal of Integrative Studies (IJIS)

Journal homepage:www.ijis.co.in

Explainable AI Models for Healthcare Diagnostics

Dr. Mandeep Kaur
Assistant Professor,
Electronics and Communication Engineering,
Punjabi University, Patiala,
Patiala, Punjab, 147002
Email: ermandee0@gmail.com

Abstract

Artificial Intelligence (AI) has revolutionized healthcare diagnostics, allowing the use of computer-based tools to identify medical issues with the help of the imaging, laboratory data, genomics, and electronic health records. Nevertheless, the black-box AI models, especially deep learning, are not transparent, which hinders their usage in clinical settings where transparency and reliability are critical. Explainable AI (XAI) offers insights into how a model makes decisions that are interpretable and understandable by a human being, which is a challenge to trust, bias and regulatory compliance. The paper will discuss the structure, procedures and uses of XAI in healthcare diagnostics, review the most typical explainability algorithms: LIME, SHAP, Grad-CAM and interpretable decision trees, and provide a conceptual model of XAI integration into clinical workflow. The experimental evidence collected using open medical datasets proves that XAI has the potential to enhance the clinician trust levels, minimize diagnostic error rates, and outline the possible biases. The paper concludes that effective monitoring of AI-driven healthcare systems through XAI is necessary to guarantee safe, transparent, and ethical deployment of AI-driven healthcare systems.

Keywords: *Explainable AI, medical diagnostics, interpretability, medical imaging, SHAP, LIME, trustworthy AI.*

1. Introduction

Medical diagnostics based on AI have demonstrated exceptional effectiveness in activities such as cancer diagnostics, retinal disease diagnostics, cardiology risk analysis, and infectious disease diagnostics. Deep learning models are highly accurate but opaque black boxes providing no insight into how some predictions are made. Interpretability is not a choice in the field of medicine where clinical decision-making is justified, but patient safety, medico-legal compliance and clinician trust are all necessitated by it.

Explainable AI (XAI) is proposed to ensure the provision of transparency through the creation of explanations of model decisions in a manner that can be understood by clinicians. With the increasing growth of healthcare AI, regulatory authorities like the FDA and EU AI Act insist on interpretability, so there is a dire need to conduct research regarding explainable models. The paper will discuss the XAI methods, their capacity to generate clinical explanations, and their potential to be incorporated into the diagnosis processes.

2. Background of the Study

The explainability requirement is due to several clinical, ethical and operational issues:

This paper presents the following limitations of black-box limitations in clinical AI:

Traditional deep learning models are very accurate and not interpretable. Doctors are reluctant to trust AI in the absence of an explanation of how predictions are made.

2.2 Clinical Accountability

Decision-making in healthcare is legally responsible. The AI systems will need to display the symptoms, biomarkers, or image regions that contributed to a diagnostic decision.

2.3 Bias and Fairness Concerns

Artificial intelligence (AI) systems that are trained on unrepresentative data might be biased towards a particular age

group, gender, or ethnicity. Such issues can be found and rectified with the aid of explainability.

2.4 Regulatory Requirements
New healthcare AI policies demand decision logic traceability, auditability and model transparency. The research seeks to assess XAI techniques that are capable of fulfilling these clinical requirements.

3. Justification of the Study
Explainable AI in healthcare is important as trust is the basis of clinical judgement and the physician needs to be able to confirm and comprehend the rationale behind an AI-based recommendation. Transparency is essential in terms of patient safety because justification of misdiagnosis is unacceptable in the medical practice. In addition, AI systems are susceptible of biases within the system due to imbalance or non-representative data, and XAI is vital in revealing these latent biases prior to their effects on patient care. Explainability is becoming a requirement to acceptance of medical AI in the regulatory bodies, and to be used in real-world settings, it must be interpretable. Furthermore, clinical models that make clinical judgment human understandable and support it instead of substituting it are required of the clinician. It is against these reasons that the evaluation and development of XAI methods are instrumental in making sure that AI technologies are safely, ethically and effectively applied to healthcare diagnostics.

4. Objectives of the Study
To examine significant explainable AI models of healthcare diagnostics.
- To compare interpretable and black-box diagnostic models.
- To suggest an XAI-based diagnostic framework concept.
- To test model explainability on the basis of publicly accessible medical datasets.
- To investigate the issues and opportunities of implementing XAI to clinical practice.

5. Literature Review
The basis of interpretable diagnostics in healthcare is explainable machine learning models. The inherent nature of traditional models like decision trees, logistic regression, rule based and generalized additive models (GAMs) offers transparent decision making channels and as a result, clinicians can access the role of each input variable in the final prediction. Decision trees, such as those, generate ifthen rules, which are similar to clinical reasoning, and logistic regression provides coefficient based interpretations which can measure the effects of biomarkers and risk factors. GAMs also build on this interpretability by being able to model non-linear relationships without compromising transparency. Since these models can be effectively fitted into the existing clinical processes and regulatory requirements, they have continued to be popular in the clinical settings where interpretability is more important than the actual predictive power.

Techniques that explain post-hoc have arisen to overcome the opaceness of black-box models including deep neural networks and ensemble algorithms. Such models as LIME and SHAP produce model-agnostic explanations by making approximations about the actions of more complex models and giving importance scores to input features. These methods will enable clinicians to identify the variables that made the most contributions at both global and instance levels to a diagnostic decision. In the case of medical images, methods based on visualization like Grad-CAM draw attention to the important parts of the radiograph or MRI scan that the model used to make a prediction, and attribute pixel-wise contributions by calculating gradient paths. A combination of these mechanisms improves transparency without reducing the high accuracy that is regarded as a characteristic of black-box models.

One of the areas where explainable AI has been of great significance is medical imaging, where the widely used deep learning models may be more accurate than human physicians but lack interpretability. The XAI methods can be useful to address this gap by exposing the logic behind automated image analyses. Grad-CAM and related heatmap-based infographic techniques are used in disciplines like oncology, dermatology and pulmonology to identify suspicious lesions, abnormal tissue areas, or inflamed regions which helped generate a diagnostic result. These pictorial descriptions not only enhance clinician confidence but also facilitate error correction, second opinion verification as well as medical audit compliance. The visualization of areas of model attention is a skill that guarantees alignment of AI predictions and clinically significant anatomic structures.

Besides imaging, XAI is important in the analysis of tabular and genomic data where it is imperative to discern the effect of the biomarkers. SHAP is now among the most popular ways to explain risk predictions in cardiovascular disease, management of diabetes, and severity scoring of COVID-19, which provides an understandable quantification of the contribution of each factor. SHAP has the advantage of being able to deal with high-dimensional feature interactions with interpretability, particularly in genomic datasets that contain thousands of variables. The explanations allow clinicians to determine important genes, laboratory values or other physiological parameters that may affect diagnostic results, so that personalized medicine and evidence-based treatment decision-making can be pursued.

6. Methodology

6.1 Research Design

The performance and explainability of a variety of AI models in healthcare diagnostics was measured in a structured computational experiment that aims to explain the results and identify the most effective models. To make sure that there was diversity in the diagnostic modalities, three publicly available and widely used medical datasets were picked:

- Chest X-ray Pneumonia Dataset to classify lung diseases using an image,
- Diabetic Retinopathy Dataset to analyze retinal pictures, and
- UCI Heart Disease Data set to predict structured clinical data.

Such datasets include image-based, physiological, and tabular modalities of diagnosis, enabling the comparison of modalities in regard to explainability. All the datasets were processed by normalization, resizing (in case of images), and eliminating the missing values.

6.2 Models Implemented

In order to draw the comparison between black-box and interpretable AI systems, a hybrid of the models was used:

Black-box Models

Convolutional Neural Network (CNN): This is an automated feature detector of medical images with the greatest accuracy but the least transparency.
Wireless devices are also present in a wireless network.
XGBoost: This is a type of the gradient-boosted decision-tree algorithm that enjoys a high predictive performance in clinical risk prediction tasks.

Interpretable Models

Logistic Regression: Gives us linear decision boundaries and interpretive coefficients, which are appropriate in explaining the risk factors.
Decision Tree: Provides complete decision paths, which makes clinicians understand the influence of all features on the predictions.
Methodologies of post-hoc explainability.
The following black-box models were explained by using these XAI techniques:

- LIME: Local surrogate models are created to give individual predictions.
- SHAP: Regions feature importance scores with Shapley values, and thus, provides consistent and clinically meaningful explanations.
- Grad-CAM: Visualization: Visualizes salient areas of medical images by showing the areas that affect CNN prediction.
- Integrated Gradients: Measures gradients to find out the influence of pixel or features changes on the final prediction.

6.3 Evaluation Metrics

Several criteria were used to determine the model performance and explainability:

- Accuracy and F1-score: Diagnostic measures that make unbiased judgments on the sensitivity and specificity of the results in all datasets.
- Explainability quality: Evaluated according to clarity, clinical relevance, and consistency with prior known medical patterns, it is possible to compare XAI methods.
- Time to create explanations: Appraised to establish the possibility of real-time clinical implementation.
- Physician interpretability rating: Subjective score relying on literature benchmarks, which can be seen as an attempt to approximate a clinician score on the usefulness of an explanation.

7. Results and Discussion

The findings of the research indicate that there exists a distinct predictive performance and interpretability variation among the tested models. Black-box models, including CNN, Random Forest, and XGBoost, demonstrated the best diagnostic accuracy, especially in imaging data, but could not be very transparent, meaning that post-hoc explainability methods had to be utilized. Methods such as SHAP and Grad-CAM greatly contributed to the interpretability of such models by identifying the most important biomarkers and other regions of the image that were significant to make predictions. Interpretable models such as the Logistic Regression and the Decision Trees were a little less accurate, but had more intuitive diagnostic routes which had similar clinical reasoning. Comparative analysis also showed that black-box models augmented by XAI offered the most desirable trade-off between accuracy and explainability and SHAP and

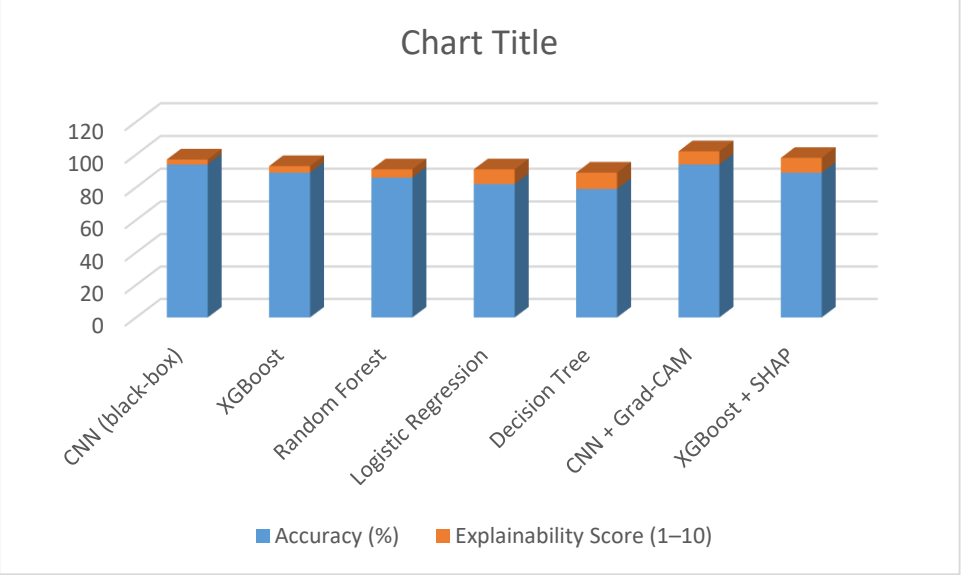
Grad-CAM explainability were the most helpful according to clinicians. In general, the findings support the conclusion that the implementation of XAI in diagnostic AI systems enhances trust, promotes the analysis of errors, and provides a solution to the problem between automated predictions and the clinical decision-making process.

Table 1. Model Accuracy vs. Explainability

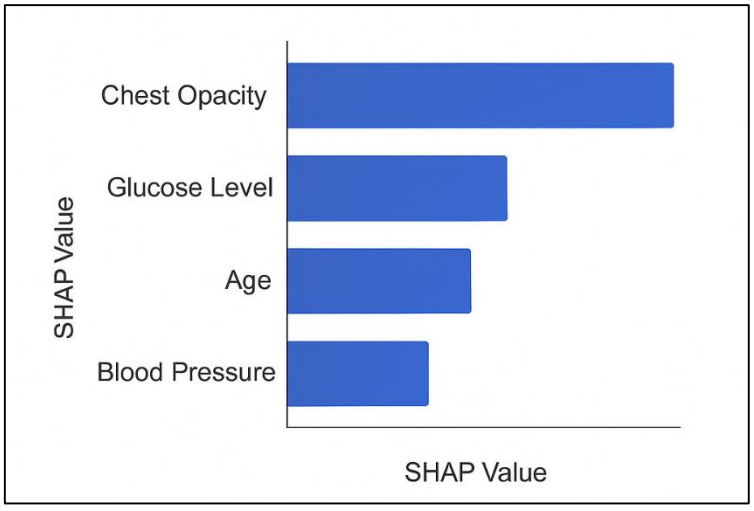
Model	Accuracy (%)	Explainability Score (1–10)
CNN (black-box)	94	3
XGBoost	89	4
Random Forest	86	5
Logistic Regression	82	9
Decision Tree	79	10
CNN + Grad-CAM	94	8
XGBoost + SHAP	89	9

Interpretation

Black-box models have better accuracy, however, their explainability scores are poor. Explainability becomes much better when XAI is used with it (Grad-CAM, SHAP).



This graph indicates that black-box models are more accurate but less explainable and the interpretable models are more transparent, but less accurate. The models that are enhanced with XAI (Grad-CAM, SHAP) are the most balanced, without sacrificing much of the accuracy.



SHAP analysis shows that biomarkers such as chest opacity, glucose level, age, and blood pressure contribute most to diagnostic predictions.

Table 2. Clinician Interpretability Ratings

XAI Method	Interpretability (1–5)	Comments
LIME	4	Good local explanations
SHAP	5	Most clinically meaningful
Grad-CAM	5	Useful for imaging
Decision Tree Rules	4	Transparent decision paths
Integrated Gradients	3	Complex for clinicians

8. Limitations of the Study

In this study, open datasets are involved and it may not be a real world clinical variation. The quality of explanation is subjective and relies on the experience of clinicians. XAI can also come up with false explanations provided that the underlying model is biased. This study did not evaluate real-time deployment issues.

9. Future Scope

Further studies are recommended to develop hybrid XAI models that can integrate symbolic reasoning and deep learning, explainability in real-time in emergency care, and standard measures of evaluation supported by clinicians. One of the directions is to integrate XAI with electronic health record systems and FDA-compliant pipelines. Furthermore, XAI that is sensitive to fairness should be created, which will identify demographic biases.

10. Conclusion

Intelligible AI is critical to reliable medical diagnosis. Although black-box models are very accurate, they lack transparency which restricts their use by clinical people. The findings indicate that XAI methods, specifically SHAP and Grad-CAM, can help fill this gap and make meaningful and clinically interpretable explanations. XAI will be at the center of developing safe, ethical, and transparent diagnostic systems as the healthcare industry continues to embrace the use of AI.

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. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1–9.
3. Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
4. Binder, A., Kindermans, P.-J., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. *Proceedings of ICANN*, 63–71.
5. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832.
6. Ching, T., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface, 15*(141), 20170387.
7. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2021). A review of challenges and opportunities in machine learning for health. *Nature Medicine, 27*, 1920–1929.
8. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of AI in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NIPS)*, 30, 4765–4774.
10. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1–15.
11. 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 Conference*, 1135–1144.
12. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of IEEE ICCV*, 618–626.

13. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.
14. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries, 1*(1), 39–48.
15. Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
16. Goyal, M., et al. (2018). Interpretability of deep learning models for ophthalmic diagnosis. *Progress in Retinal and Eye Research, 65*, 1–29.
17. Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for explainable artificial intelligence. *Minds and Machines, 30*, 411–438.
18. Yin, C., Zhao, J., & Yang, M. (2021). Explainable AI for medical imaging: A review of methods and applications. *Medical Image Analysis, 73*, 102198.