International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Explainable AI for Healthcare Diagnostics

Khushi Khandelwal

Abstract

The integration of Artificial Intelligence (AI) in healthcare diagnostics has revolutionized the field by enabling faster, more accurate, and data-driven decision-making. However, the complexity and opacity of many AI models, particularly deep learning algorithms, pose significant challenges to trust, accountability, and clinical adoption. Explainable AI (XAI) emerges as a critical solution to bridge this gap, offering transparency and interpretability to AI-driven healthcare systems. This paper explores the significance of XAI in healthcare diagnostics, examining various methods such as SHAP, LIME, and attention mechanisms that enhance model interpretability without compromising performance. We analyze real-world case studies and develop a working diagnostic model with explainable outputs, demonstrating how XAI can improve clinician trust, patient safety, and regulatory compliance. The study concludes with insights into the current challenges and future directions of implementing XAI in clinical settings.

1. INTRODUCTION

Introduction:

Artificial Intelligence (AI) is transforming healthcare by enabling faster and more accurate diagnostics. However, many AI models function as "black boxes," offering little insight into how decisions are made. In high-stakes fields like healthcare, this lack of transparency can hinder trust and adoption. Explainable AI (XAI) addresses this issue by making AI predictions more understandable to clinicians and patients. This research explores the role of XAI in healthcare diagnostics, aiming to improve transparency, trust, and clinical decision-making through interpretable AI models.

1.1 Literature Reivew

The integration of Artificial Intelligence (AI) in healthcare diagnostics has revolutionized the field by enabling faster, more accurate, and data-driven decision-making. However, the complexity and opacity of many AI models, particularly deep learning algorithms, pose significant challenges to trust, accountability, and clinical adoption. Explainable AI (XAI) emerges as a critical solution to bridge this gap, offering transparency and interpretability to AI-driven healthcare systems. This paper explores the significance of XAI in healthcare diagnostics, examining various methods such as SHAP, LIME, and attention mechanisms that enhance model interpretability without compromising performance. We analyze real-world case studies and develop a working diagnostic model with explainable outputs, demonstrating how XAI can improve clinician trust, patient safety, and regulatory compliance. The study concludes with insights into the current challenges and future directions of implementing XAI in clinical settings.

1.1 Data Collection

For this research, medical diagnostic datasets were collected from publicly available sources such as Kaggle and government health repositories. The datasets included labeled patient records and medical imaging data related to diseases like diabetes, heart conditions, and pneumonia. All datasets were pre-processed to remove missing values, normalize features, and ensure patient anonymity. These clean and structured datasets were then used to train machine learning models and apply explainability techniques like SHAP and LIME.

2. Applications of Deepfake Detection:

1.Deepfake Detection Training: Used to train AI models and forensic tools to identify and prevent the spread of manipulated or fake media.

2.Research and Education: Helps in studying the ethical, psychological, and technological aspects of deepfakes in academic and research settings.

3. Methodology

Public medical datasets (e.g., UCI Heart Disease, Chest X-ray for Pneumonia) were collected from open-source repositories. Data preprocessing included handling missing values, normalization, and encoding categorical variables.

3.2 Model Development

A Random Forest and a Convolutional Neural Network (CNN) were trained on structured and image-based datasets, respectively, to predict disease outcomes.

3.3 Explainability Techniques

- SHAP was applied to structured data to interpret feature contributions.
- LIME was used to explain individual predictions in both models.
- Saliency Maps were used in CNN to highlight image regions influencing predictions.

4. Results and Discussion

The application of SHAP revealed key contributing factors in heart disease predictions such as cholesterol levels and resting ECG results. LIME provided instance-specific explanations that aligned well with clinical knowledge. Saliency maps in the pneumonia detection model correctly highlighted infected lung regions, supporting the reliability of CNN predictions.

Clinicians found these explanations intuitive and useful, suggesting improved trust in AI-assisted diagnoses. However, the study also revealed limitations in consistency and performance when explanations varied across similar cases, indicating the need for further refinement in XAI techniques.Countries around the world are incorporating gamification in diverse ways. For instance, Finland integrates playful learning strategies in elementary education, while the U.S. uses advanced learning management systems with gamification features in higher

5. Applications and Case Studies

Explainable AI is becoming increasingly vital across various medical domains, particularly in diagnostics, risk assessment, and personalized medicine. The following applications and case studies illustrate how XAI contributes to enhancing clinical trust, model validation, and real-world implementation

Disease Diagnosis and Risk Prediction

One of the most impactful applications of XAI is in disease diagnosis and risk prediction, where transparency is essential. For example:

- Cardiovascular Risk Prediction: Models like Random Forest or Gradient Boosting are used to predict heart disease based on features like age, cholesterol, and blood pressure. SHAP values are applied to rank feature importance, helping physicians understand which factors are contributing most to individual risk predictions.
- Diabetes Onset Detection: Logistic regression and tree-based models trained on datasets like Pima Indians Diabetes Dataset are made explainable using LIME. The model highlights key indicators such as glucose level and BMI, which helps in early diagnosis and lifestyle guidance.

Medical Imaging and Radiology

In imaging-based diagnostics, explainable AI plays a key role in validating and visualizing deep learning model outputs:

- Case Study: Pneumonia Detection from Chest X-rays

A CNN trained on the NIH Chest X-ray dataset was used to detect pneumonia. To make the model explainable, saliency maps and Grad-CAM were applied to highlight the specific regions in the lungs that led to a positive diagnosis. Radiologists could confirm that the model was focusing on medically relevant areas, improving their confidence in its use.

Key Benefits Observed

- Improved trust and adoption of AI by clinicians.
- Easier regulatory approval due to transparency.
- Enhanced error detection, where clinicians could catch incorrect model assumptions.
- Support for human-in-the-loop systems that combine AI efficiency with human expertise..

6. Conclusion

This research demonstrates the critical role of Explainable AI in enhancing transparency, trust, and safety in healthcare diagnostics. By applying SHAP, LIME, and saliency maps to diagnostic models, we showcased how XAI enables clinicians to better understand and trust AI recommendations. Future work will focus on integrating these techniques into real-time clinical systems and ensuring compliance with healthcare regulations.

References

1.Caruana, R., et al. (2015). Intelligible models for healthcare. Proceedings of the 21th ACM SIGKDD International Conference.

2.Ribeiro, M. T., et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD.

3. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions.