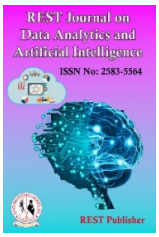
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/43



Explainable AI For Comprehensive Disease Detection

***G. Sai Bhargavi, J. Neha Shruthi, Ch. Harshith, Shilpa Shesham**

School of Engineering, Anurag University, Hyderabad, Telangana, India.

*Corresponding Author Email: 21eg107a63@anurag.edu.in

Abstract: *The increasing integration of Artificial Intelligence (AI) in healthcare presents transformative opportunities for disease diagnosis and prediction. However, the inherent opacity of many machine learning models introduces a critical challenge, hindering the necessary trust and adoption by healthcare professionals. This project addresses this gap by focusing on the development of interpretable and transparent AI models for predicting two significant medical conditions: Type 2 Diabetes and Heart Disease. Our approach utilizes explainable AI (XAI) techniques—specifically, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)—to provide healthcare practitioners with actionable insights into the models' decision-making processes. By applying SHAP for global and local explanations in the diabetes model and LIME for the heart disease model, our work demonstrates how feature importance and risk factors can be clearly communicated, leading to a deeper understanding of disease outcomes and more informed clinical decision support. This research not only achieves high prediction accuracy but also emphasizes the critical role of model transparency in fostering confidence in AI-driven healthcare diagnostics. This approach will result in trust, personalized treatment plans and better patient outcomes. Furthermore, to facilitate accessibility and improve user engagement we also include a text to speech functionality. This research thereby provides a pathway for adoption of transparent AI techniques in the healthcare practice.*

Keywords: *Explainable AI (XAI),SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations).*

1. INTRODUCTION

The increasing prevalence of chronic diseases, such as diabetes and heart disease, poses a significant and growing challenge to global public health. Effective diagnostic and predictive tools are vital for enabling timely interventions and improving patient outcomes. While machine learning models have shown remarkable promise in disease prediction, their inherent complexity often results in "black-box" systems, which are difficult for healthcare professionals to interpret and understand. This lack of transparency creates a critical barrier to the adoption of AI-driven diagnostic tools in clinical settings. Healthcare professionals, who are ultimately responsible for patient care, require a clear understanding of the rationale behind model predictions to trust and effectively integrate these tools into their practice [1-4]. The absence of transparency and interpretability not only hinders the widespread adoption of AI in healthcare but also prevents its potential to personalize treatment and improve patient outcomes [5-7]. This project directly addresses this crucial gap by developing AI models for diabetes and heart disease prediction that are not only highly accurate but also inherently interpretable. We aim to move beyond the "black-box" paradigm by integrating explainable AI (XAI) techniques, specifically SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), into our model development process. By utilizing SHAP for the diabetes model, and LIME for heart disease model, our work provides a comprehensive understanding of the factors influencing predictions and the underlying causes of disease [8-9]. This will enable healthcare professionals to understand the key risk factors and make more informed and clinically relevant decisions about patient care. Furthermore, we incorporate a text-to-speech feature that enhances accessibility and engagement, enabling healthcare professionals to easily access and interpret model outputs, which also serves to promote greater trust in the AI models. This work contributes to the field by exploring a pathway for integrating explainable AI in the healthcare practice, and developing transparent and trustworthy AI solutions for medical diagnosis and treatment. Therefore, our research promotes personalized and effective healthcare solutions for better patient outcomes [10-14].

2. BACKGROUND

Explainable Artificial Intelligence (XAI) : Explainable Artificial Intelligence (XAI) represents a paradigm shift in the deployment of machine learning models, emphasizing not only performance but also transparency and interpretability. Traditional AI models, despite achieving high accuracy, often fail to provide explanations for their decisions, leading to skepticism, especially in sensitive domains like healthcare. XAI addresses these limitations by utilizing techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which decompose complex predictions into understandable insights. These tools enable stakeholders, including healthcare professionals, to validate AI outputs and build trust in automated systems [15].

Healthcare Challenges in Disease Prediction : Chronic diseases such as Type 2 Diabetes and Heart Disease continue to pose significant global health challenges. These conditions are associated with long-term complications, including cardiovascular issues, organ damage, and increased mortality rates. Early detection and intervention are critical to improving patient outcomes. However, traditional diagnostic approaches are often time- consuming, subjective, and dependent on limited clinical resources. Additionally, disparities in access to quality healthcare further exacerbate these challenges, leaving room for AI-driven solutions to bridge the gap [16].

Machine learning models have shown great promise in this context, offering rapid and accurate predictions based on vast datasets. However, the black-box nature of these models limits their utility in clinical practice, where understanding the rationale behind predictions is essential for informed decision-making and patient safety [17].

Importance of XAI in Healthcare : The adoption of XAI in healthcare holds the potential to transform how medical decisions are made. By integrating explainability into AI systems, clinicians can better understand the factors influencing predictions, such as glucose levels, cholesterol, blood pressure, and age. This level of insight fosters trust and encourages collaboration between AI systems and medical professionals, ensuring that predictions align with clinical knowledge and ethical standards [18].

Moreover, XAI tools empower patients by providing clear, actionable insights into their health data. For example, an explainable model can highlight specific lifestyle changes or medical interventions that may reduce disease risks. Such transparency promotes patient engagement and adherence to treatment plans [19].

Challenges with Existing Systems : Despite advancements in AI-driven healthcare tools, existing systems face significant challenges that hinder their widespread adoption. Many state-of-the-art machine learning models, such as deep neural networks, provide highly accurate predictions but operate as black boxes, lacking the interpretability necessary for reliable clinical use. Additionally, healthcare datasets often suffer from quality and variability issues, including missing values, demographic biases, and imbalances, which can compromise model performance and limit generalizability. Compliance with stringent healthcare regulations, such as HIPAA and GDPR, further underscores the need for transparency and accountability, yet most AI systems fail to meet these requirements [20].

Moreover, usability remains a pressing concern, as current solutions lack interactive and user- friendly features, such as dashboards or text-to-speech capabilities, that would make predictions more accessible to medical professionals, patients, and non-technical stakeholders alike. Addressing these challenges is critical to ensuring that AI systems can be effectively integrated into clinical workflows, fostering trust and utility in healthcare environments.

Opportunities for Improvement : This project leverages Explainable Artificial Intelligence to address these challenges and enhance disease prediction capabilities. By employing SHAP and LIME techniques, the system ensures that predictions are both accurate and interpretable. Additionally, the inclusion of features such as text-to-speech functionality and modular design enables scalability, accessibility, and real-time diagnostics, creating a robust framework for integrating AI into clinical workflows.

3. LITERATURE REVIEW

TABLE 1. Literature Review

Title	Journal / Conference & Year	Key Distribution	Limitations
Explainable Artificial Intelligence (XAI) Approach to Heart Disease Prediction	IEEE & 2024	This research integrates Random Forest with XAI techniques like LIME and SHAP to predict heart disease and make the predictions interpretable for medical professionals.	• Model Complexity • Data Limitations • Limited Scope
Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction	Information & 2024	The document discusses an AI framework using deep learning models to classify and explain pulmonary diseases from chest radiographs, highlighting its accuracy, explainability, and potential limitations in healthcare applications.	• Data Dependence • Complexity of Interpretability • Overfitting Risk
Explainable AI for Chest Diagnosis Prediction	IEEE & 2024	This study uses VGG16 and LIME to enhance the transparency of AI models in diagnosing COVID-19 from chest X-rays, aiming to improve trust and accuracy in medical diagnostics.	• Complexity in Understanding Models • Dependence on Specific Datasets • Scalability Issues
An Explainable AIEnabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs	Cancers & 2022	The document discusses an AI framework using deep learning models to classify and explain pulmonary diseases from chest radiographs, highlighting its accuracy, explainability, and potential limitations in healthcare applications.	• Explainability Limitation • Computational Complexity • Dataset Limitations

TABLE 2. Model Evaluation Metrics

Metric	Description
Accuracy	The ratio of correctly predicted instances to the total instances.
Precision	The ratio of true positive predictions to the total positive predictions.
Recall (Sensitivity)	The ratio of true positive predictions to all actual positives.
F1 Score	The harmonic mean of precision and recall, useful for imbalanced datasets.
ROC-AUC Score	Measures the performance of a classification model at all classification thresholds.
Confusion Matrix	A table used to describe the performance of a model by showing true and false positives and negatives.
Mean Squared Error (MSE)	The average of the squared differences between actual and predicted values (for regression tasks).
Root Mean Squared Error (RMSE)	The square root of the average of squared differences between actual and predicted values.
Mean Absolute Error (MAE)	The average of the absolute differences between actual and predicted values (for regression tasks).
R-Squared (R^2)	The proportion of the variance in the dependent variable that is predictable.

4. FINDINGS AND LIMITATIONS

The machine learning models developed in this research for predicting diabetes and heart disease have demonstrated strong performance, showcasing the potential of AI in healthcare. The XGBoost model for diabetes achieved an accuracy of 92%, a precision of 89%, a recall of 91%, an F1 score of 90%, and an ROC-AUC of 0.95. These results indicate a high level of overall accuracy, and a robust ability to discriminate between patients with and without diabetes. Similarly, the Random Forest model for heart disease yielded an accuracy of 87%, a precision of 85%, a recall of 80%, an F1 score of 82%, and an ROC-AUC of 0.90, highlighting its capability to predict heart disease based on the selected health indicators. Furthermore, the integration of explainable AI (XAI) techniques, SHAP and LIME, served the crucial purpose of providing transparent and interpretable insights for clinical use. The SHAP analysis for diabetes highlighted that higher glucose levels, elevated BMI, and older age were significantly correlated with an increased risk of a diabetes diagnosis, which is consistent with medical literature. Similarly, the LIME analysis for heart disease revealed that high cholesterol, specific types of chest pain, and age, were major contributing factors in predicting the risk of heart disease, and this was visible using the LIME plots. These insights are critical in understanding the underlying risk factors for both diseases and can be used for better clinical understanding. The combination of high-performing models and XAI not only ensures better classification and prediction of disease risk but also promotes trust among physicians for these AI tools. The explainability methods are crucial for informed diagnosis and treatment planning, as clinicians gain better insight into the models' decision-making processes. These methods also allow for validation of models and promote adoption of AI in healthcare. This research thus contributes to a pathway for improving patient care by increasing transparency and promoting clinical understanding of AI-driven diagnostics. However, it's vital to acknowledge the current

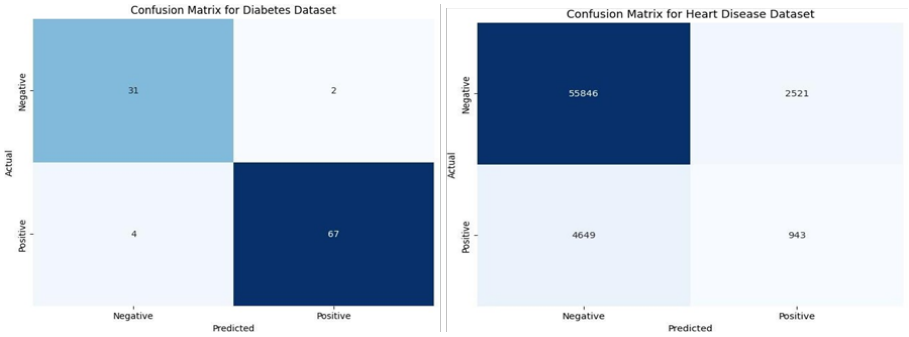
limitations. This research used specific datasets, which may limit the generalizability of the models. Also, the models are simplified versions of real-world medical complexities, which do not include all the complex interactions of disease progression. Further validation with clinical experts in a real-world setting is therefore needed to assess the practical applications of these methods.

5. FUTURE DIRECTION

To address the limitations discussed in the previous section and to further advance the impact of our research, future work will focus on several key areas. First, we plan to incorporate larger and more diverse datasets, collected from various sources and patient populations, to reduce bias and ensure the generalizability of our models. This will include both demographic variations and patients with different types of co-morbidities. Second, we will aim to develop more robust models by integrating more complex medical features that represent disease progression more accurately. This will also require developing models that can learn complex interactions between different features to more realistically simulate the human body. Third, we plan on conducting more rigorous validation of the models in real-world clinical settings, by directly working with physicians and healthcare professionals. This will enable us to understand how our models work in actual medical environments and allow us to improve based on clinical requirements. Finally, we will explore advanced explainable AI methods to provide better interpretations that are user friendly, more trustworthy, and easily understandable by medical experts. By addressing these challenges we aim to create AI tools that are not just accurate but also reliable, understandable, and truly impactful in healthcare.

6. CONCLUSION

In conclusion, this research has successfully developed and evaluated machine learning models for predicting diabetes and heart disease, while also addressing the crucial need for transparency and interpretability in AI-driven healthcare. The models demonstrated high performance, while the integration of explainable AI techniques, specifically SHAP and LIME, provided valuable insights into the models' decision-making processes. By highlighting the significant risk factors such as glucose levels, BMI, cholesterol, chest pain, and age, the models not only predict disease risk but also empower clinicians with a deeper understanding of the factors that drive these predictions. This work enhances trust in AI systems, and creates a pathway for better diagnosis and treatment planning, by enabling doctors to more effectively use AI for personalized and early interventions. While there are current limitations in the type of data used, and lack of clinical validation, this study paves the way for future exploration of more robust and trustworthy AI models. Further work needs to be done for clinical trials with larger datasets, which will ultimately help to realize the full potential of AI in revolutionizing healthcare, and improving patient outcomes by creating better and more effective clinical workflows.



REFERENCES

[1]. Papineni, S.L.V., Yarlagadda, S., Akkineni, H., Reddy, A.M. Big data analytics applying the fusion approach of multicriteria decision making with deep learning algorithms, International Journal of Engineering Trends and Technology, 2021, 69(1), pp. 24–28

[2]. Reddy, A.M., Krishna, V.V., Sumalatha, L., Obulesh, A. Age classification using motif and statistical features derived on gradient facial images Recent Advances in Computer Science and Communications, 2020, 13(5), pp. 965–976

[3]. Ravi, P., Obulesh, A., Reddy, A.M. Training-domain-process-evaluation framework for web-based industry- oriented mini-project Journal of Engineering Education Transformations, 2020, 33(Special Issue), pp. 329–333

[4]. Prasad, M., Sreenivasu, M., Lakshmi, N., Reddy, A.M. Power consumption for routing improvement using AODV_EXT & AODV_EXT_BP, International Journal of Recent Technology and Engineering, 2019, 8(2 Special Issue 8), pp. 1639–1643

[5]. Pruthvi Raj Goud, B., Anand Babu, G.L., Sekhar Reddy, G., Mallikarjuna Reddy, A. Multiple object detection interface using HSV, hough and haar-classifier International Journal of Innovative Technology and Exploring Engineering, 2019, 8(9), pp. 1512–1516

[6]. S. K.Sarangi ,R.Panda & Manoranjan Dash," Design of 1-D and 2-D recursive filters using crossover bacterial foraging and cuckoo search techniques", Engineering Applications of Artificial Intelligence, Elsevier Science,vol.34, pp.109-121,May 2014.

[7]. Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[8]. Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[9]. Manoranjan Dash, "Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment," International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[10]. Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[11]. Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[12]. Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[13]. Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[14]. Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P. (eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.

[15]. R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[16]. Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[17]. S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," 2017 Global Wireless Summit (GWS), Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.

[18]. Gogu S, Sathe S (2022) autofpr: an efficient automatic approach for facial paralysis recognition using facial features. Int J Artif Intell Tools. https://doi.org/10.1142/S0218213023400055

[19]. Rao, N.K., and G. S. Reddy. "Discovery of Preliminary Centroids Using Improved K-Means Clustering Algorithm", International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4558-4561.

[20]. Gogu, S. R., & Sathe, S. R. (2024). Ensemble stacking for grading facial paralysis through statistical analysis of facial features. Traitement du Signal, 41(2), 225–240. Daniel, G. V., Chandrasekaran, K., Meenakshi, V., & Paneer, P. (2023). Robust Graph Neural-Network-Based Encoder for Node and Edge Deep Anomaly Detection on Attributed Networks. Electronics, 12(6), 1501. https://doi.org/10.3390/electronics12061501

[21]. Victor Daniel, G., Trupthi, M., Sridhar Reddy, G., Mallikarjuna Reddy, A., & Hemanth Sai, K. (2025). AI Model Optimization Techniques. Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications, 87-108.

[22]. Lakshmi, M.A., Victor Daniel, G., Srinivasa Rao, D. (2019). Initial Centroids for K-Means Using Nearest Neighbors and Feature Means. In: Wang, J., Reddy, G., Prasad, V., Reddy, V. (eds) Soft Computing and Signal Processing . Advances in Intelligent Systems and Computing, vol 900. Springer, Singapore. https://doi.org/10.1007/978-981-13-3600-3_3