Responsible AI: Explainable Artificial Intelligence for Heart Disease Prediction using LIME

Devansh Agarwal[1]

[1]Student, Department of Computer Science & Engineering, Birla Institute of Technology, Jaipur, India

Agarwaldevansh88@gmail.com

Abstract. Artificial intelligence has increasingly been adopted in healthcare for predictive diagnosis and clinical decision support. However, many high-performing machine learning models operate as opaque "black-box" systems, limiting their transparency and raising concerns regarding trust, accountability, and responsible deployment in critical domains such as medicine. Responsible AI principles emphasize the importance of interpretability, fairness, and transparency to ensure that automated decision-making systems can be understood and validated by human experts. In this study, an explainable machine learning framework is proposed for heart disease prediction using clinical data from the UCI Heart Disease dataset. The dataset consists of 920 patient records and multiple clinical attributes related to cardiovascular health. Three classification algorithms: Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) were implemented and evaluated using performance metrics including accuracy, ROC-AUC, and cross-validation. Experimental results indicate that the XGBoost model achieved the best predictive performance with an accuracy of approximately 85.3%. To enhance transparency and align with responsible AI practices, the Local Interpretable Model-Agnostic Explanations (LIME) technique was applied to generate interpretable explanations for individual predictions. Global feature importance and local LIME explanations were analyzed to identify clinically relevant attributes influencing heart disease prediction. The results demonstrate that integrating explainable AI methods with machine learning models improves transparency and supports the development of trustworthy AI-driven healthcare systems capable of assisting clinicians in informed decision-making.

Keywords: Explainable Artificial Intelligence, Responsible AI, Heart Disease Prediction, LIME

1. INTRODUCTION

Cardiovascular diseases remain one of the leading causes of mortality worldwide, accounting for a significant proportion of global deaths each year. Early detection and accurate diagnosis of heart disease are therefore critical for improving patient outcomes and reducing mortality rates. With the rapid growth of healthcare data and advancements in computational methods, machine learning techniques have emerged as powerful tools for assisting medical professionals in disease prediction and decision support. These techniques are capable of identifying complex patterns in

clinical datasets and generating predictive models that can aid in the early detection of cardiovascular conditions.

In recent years, several machine learning algorithms have been applied to heart disease prediction, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting models. These algorithms can analyze patient health indicators such as age, cholesterol levels, chest pain type, electrocardiographic results, and exercise-induced angina to estimate the likelihood of cardiovascular disease. Ensemble learning techniques, particularly gradient boosting algorithms such as Extreme Gradient Boosting (XGBoost), have demonstrated strong predictive performance due to their ability to capture nonlinear relationships within medical datasets.

Despite the predictive capabilities of machine learning models, many advanced algorithms operate as black-box systems, meaning that their internal decision-making processes are not easily interpretable. This lack of transparency presents a major challenge in healthcare applications, where clinicians must understand and justify diagnostic decisions. Without interpretable explanations, medical practitioners may hesitate to rely on automated prediction systems, especially in critical clinical environments where accountability and trust are essential.

To address this limitation, the field of Explainable Artificial Intelligence (XAI) has emerged as an important research area aimed at improving the transparency and interpretability of machine learning models. Explainable AI techniques provide insights into how models make predictions by identifying the most influential features contributing to a decision. Such explanations help bridge the gap between complex computational models and human understanding, enabling clinicians to evaluate whether the model's reasoning aligns with established medical knowledge.

One widely used XAI technique is Local Interpretable Model-Agnostic Explanations (LIME), which explains individual predictions by approximating complex machine learning models with simpler interpretable models in the local region around a data instance. Rather than attempting to interpret the entire model globally, LIME focuses on explaining specific predictions, making it particularly suitable for healthcare applications where patient-level explanations are required.

In this study, we propose an explainable machine learning framework for heart disease prediction using the UCI Heart Disease dataset, which contains 920 patient records and 16 clinical attributes. The dataset includes important medical features such as age, sex, chest pain type, cholesterol levels, resting blood pressure, exercise-induced angina, and electrocardiographic measurements. Three machine learning models: Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) are trained and evaluated to determine the most effective predictive model.

Experimental evaluation shows that XGBoost achieved the highest prediction accuracy of approximately 85.3%, outperforming Logistic Regression and Random Forest models. The predictive performance of the models was further evaluated using metrics such as ROC-AUC

score and cross-validation, ensuring the reliability of the results. To enhance interpretability, LIME is applied to generate local explanations for individual predictions, highlighting the most influential clinical features contributing to heart disease risk. Additionally, a comparison between global feature importance derived from the trained model and local explanations produced by LIME is performed to provide a comprehensive understanding of model behavior.

The proposed approach demonstrates that combining machine learning prediction models with explainable AI techniques can significantly improve transparency and trust in AI-based healthcare systems. By providing interpretable insights into model decisions, the framework can assist healthcare professionals in validating predictions and integrating AI-driven tools into clinical decision-making processes.

2. LITERATURE REVIEW

The rapid growth of healthcare data and advances in computational techniques has significantly expanded the use of machine learning methods in medical diagnosis and disease prediction. Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early detection and risk prediction a critical focus in medical research. Machine learning models have increasingly been applied to analyze clinical datasets and identify patterns associated with cardiovascular risk. These models are capable of processing multiple physiological and clinical attributes simultaneously, allowing them to detect complex relationships among variables that may not be easily identified through traditional statistical approaches.

Early research in heart disease prediction primarily relied on conventional statistical models such as Logistic Regression due to their simplicity and interpretability. Logistic Regression remains widely used in medical studies because it provides probabilistic predictions and allows researchers to understand the influence of individual features on disease outcomes. However, as healthcare datasets have grown in complexity and dimensionality, traditional statistical methods have often been supplemented or replaced by more advanced machine learning algorithms capable of modeling nonlinear relationships within data.

In recent years, ensemble learning techniques have gained significant attention for cardiovascular disease prediction. Algorithms such as Random Forest and gradient boosting models have demonstrated strong performance in structured medical datasets. Random Forest improves predictive stability by aggregating the outputs of multiple decision trees, thereby reducing overfitting and variance. Gradient boosting methods, including Extreme Gradient Boosting (XGBoost), further enhance predictive capability by sequentially correcting errors made by previous models. These algorithms have shown particularly strong performance in healthcare analytics due to their ability to capture complex interactions among clinical variables.

Despite the high predictive accuracy achieved by these machine learning techniques, the interpretability of many models remains a major challenge. Advanced algorithms such as ensemble methods and deep learning architectures often operate as black-box systems, meaning

that the reasoning behind their predictions is not easily interpretable. In healthcare applications, this lack of transparency can limit the practical adoption of artificial intelligence systems, as medical professionals require clear explanations to validate predictions and ensure that automated decisions align with clinical knowledge.

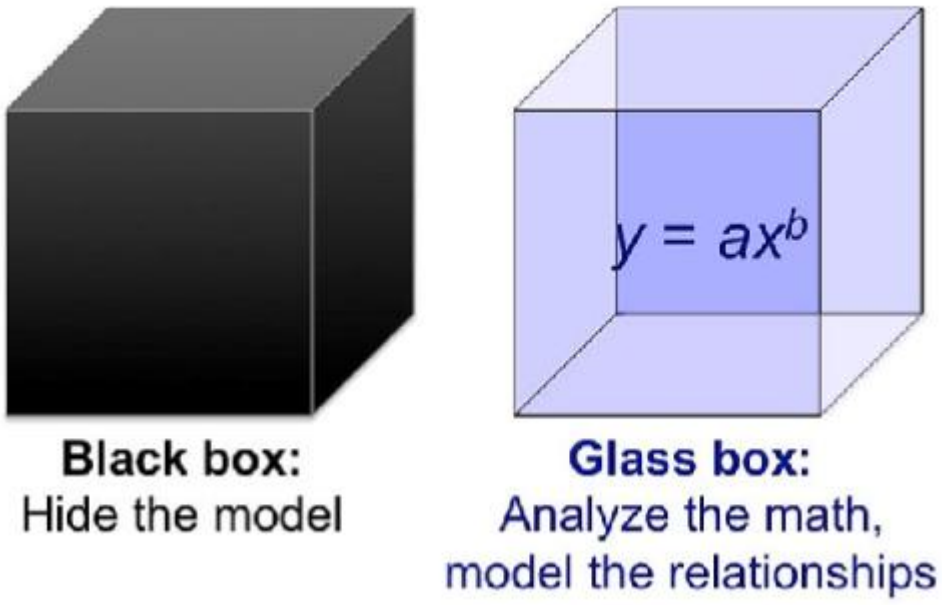

Figure 1: Black Box to Glass Box

LIME provides local explanations by approximating the behaviour of a model around a specific instance using a simpler interpretable model. By generating perturbed samples around a data point and analyzing the resulting predictions, LIME identifies the features that contribute most significantly to the prediction outcome. This approach allows researchers and practitioners to understand the factors influencing individual predictions without requiring direct access to the internal structure of the model.

Recent studies in healthcare analytics have increasingly emphasized the importance of combining predictive modeling with explainability techniques to improve both accuracy and interpretability in medical decision support systems. The integration of machine learning algorithms with explainable AI frameworks enables researchers to evaluate not only the predictive performance of models but also the relevance of clinical features influencing the predictions. Such approaches help ensure that the model's behaviour aligns with established medical knowledge and provide meaningful insights for healthcare practitioners.

Furthermore, recent developments in explainable AI for cardiovascular disease prediction highlight the importance of integrating interpretable machine learning frameworks into clinical applications. Research in this area has demonstrated that incorporating explainability methods

alongside predictive models can improve the reliability and transparency of automated diagnostic systems, thereby supporting the adoption of artificial intelligence in healthcare environments.

Although significant progress has been made in developing predictive models for heart disease, many studies focus primarily on improving prediction accuracy while providing limited analysis of model interpretability. There remains a need for research that simultaneously evaluates multiple machine learning algorithms and integrates explainability techniques to provide both global insights into feature importance and local explanations for individual predictions. Addressing this gap, the present study investigates the application of multiple machine learning models for heart disease prediction and incorporates explainable AI techniques to analyze both global and local feature contributions in the prediction process.

3. METHODOLOGY

3.1 Overview of the Proposed Framework

The proposed framework integrates machine learning prediction with explainable artificial intelligence techniques to develop a transparent heart disease prediction system. The methodology consists of multiple stages, including data preprocessing, model training, performance evaluation, and interpretability analysis. Multiple machine learning models are trained and evaluated to identify the most effective classifier for predicting heart disease. The best-performing model is then interpreted using explainable AI techniques to provide insights into the factors influencing the prediction outcomes.

The overall workflow of the proposed framework is illustrated as follows:

Dataset Collection → Data Preprocessing → Model Training → Model Evaluation → Best Model Selection → Global Feature Importance → LIME Explanation

This pipeline ensures that the predictive model achieves both high accuracy and interpretability, which are essential for healthcare applications.

3.2 Dataset Description

The study utilizes the UCI Heart Disease dataset, which is widely used for cardiovascular risk prediction research. The dataset contains 920 patient records with 16 attributes, including demographic and clinical features relevant to cardiovascular diagnosis.

The dataset includes information collected from multiple medical institutions and contains attributes related to patient health indicators such as age, gender, chest pain type, cholesterol levels, blood pressure, and electrocardiographic measurements. The dataset also contains a target variable that indicates the presence and severity of heart disease.

Attribute	Description
age	Age of the patient
sex	Gender of the patient
dataset	Source hospital dataset
cp	Chest pain type
trestbps	Resting blood pressure
chol	Serum cholesterol level
fbs	Fasting blood sugar
restecg	Resting ECG results
thalch	Maximum heart rate achieved
exang	Exercise-induced angina
oldpeak	ST depression induced by exercise
slope	Slope of the ST segment
ca	Number of major vessels colored by fluoroscopy
thal	Thalassemia condition
num	Diagnosis of heart disease

Table 1: a summary of the dataset attributes.

The target variable (num) originally contains multiple classes representing the severity of heart disease. For the purpose of this study, the problem is converted into a binary classification task, where:

- 0 indicates absence of heart disease
- 1 indicates presence of heart disease

This transformation simplifies the prediction task while maintaining the clinical relevance of the dataset.

3.3 Data Preprocessing

Data preprocessing is an essential step to ensure that the dataset is suitable for machine learning algorithms. Several preprocessing operations were performed to prepare the dataset for training.

First, the identifier column was removed because it does not contribute to the predictive process. Next, categorical variables such as gender, chest pain type, ECG results, and thalassemia condition were converted into numerical representations using one-hot encoding. This transformation allows machine learning algorithms to interpret categorical attributes as numerical inputs.

The dataset also contains missing values in several clinical attributes, which were handled using median imputation. Median imputation was chosen because it is robust to outliers and preserves the distribution of numerical variables.

After handling missing values and encoding categorical variables, feature scaling was performed using standardization. Standard scaling ensures that all features have comparable ranges and prevents models from being biased toward variables with larger magnitudes.

Finally, the dataset was divided into training and testing subsets using an 80–20 split, ensuring that the class distribution remained balanced through stratified sampling.

3.4 Machine Learning Models

To evaluate the predictive performance of different algorithms, three machine learning models were implemented and compared.

Logistic Regression: Logistic Regression is a widely used statistical classification technique for binary prediction problems. It estimates the probability of a class using a logistic function and provides interpretable coefficients representing the influence of input variables on the predicted outcome. Logistic Regression serves as a baseline model for evaluating more complex algorithms.

Random Forest: Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to produce the final output. By combining several trees, Random Forest improves prediction accuracy and reduces overfitting. The algorithm is particularly effective for structured datasets and can capture nonlinear relationships between features.

Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting algorithm designed to improve predictive performance and computational efficiency. The algorithm sequentially builds decision trees, where each new tree attempts to correct the errors made by previous trees. XGBoost incorporates regularization techniques that

help prevent overfitting and improve generalization. Due to its strong performance on structured datasets, XGBoost has become widely used in predictive analytics and healthcare research.

3.5 Model Evaluation

The predictive performance of each model was evaluated using multiple evaluation metrics to ensure reliability and robustness.

The following metrics were used:

- Accuracy: Measures the proportion of correctly classified instances.
- ROC-AUC Score: Evaluates the model's ability to distinguish between positive and negative classes.
- Cross-validation Score: Provides an estimate of the model's performance on unseen data. Cross-validation was performed using a five-fold validation strategy, which divides the dataset into multiple subsets to evaluate model performance across different training and testing partitions.

Based on the experimental results, XGBoost achieved the highest prediction accuracy (approximately 85%), outperforming both Logistic Regression and Random Forest. Consequently, XGBoost was selected as the best-performing model for further interpretability analysis.

3.6 Global Feature Importance

To understand the overall behavior of the predictive model, global feature importance analysis was performed using the trained XGBoost model. Feature importance measures the contribution of each feature to the model's predictive performance across the entire dataset.

The analysis identified several important clinical attributes influencing heart disease prediction, including:

- Chest pain type (asymptomatic)
- Exercise-induced angina
- Chest pain type (atypical angina)
- Dataset source indicators
- Thalassemia condition
- Gender
- ST segment slope
- Number of major vessels

These features correspond to clinically relevant indicators associated with cardiovascular risk, highlighting the consistency between the machine learning model and established medical knowledge.

3.7 Local Interpretability using LIME

While global feature importance provides insights into overall model behavior, it does not explain individual predictions. To address this limitation, Local Interpretable Model-Agnostic Explanations (LIME) was applied to generate explanations for specific patient predictions.

LIME works by generating perturbed samples around a data instance and analyzing how these changes influence the model's predictions. A simple interpretable model is then fitted locally to approximate the behavior of the complex model.

This process produces a ranked list of features that contributed most strongly to the prediction for a specific patient. For example, LIME explanations identified factors such as chest pain type, cholesterol levels, exercise-induced angina, and age-related attributes as influential features for certain predictions.

By comparing global feature importance with LIME-based local explanations, the study provides a comprehensive understanding of both overall model behavior and individual decision outcomes.

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

The experiments were conducted using the UCI Heart Disease dataset containing 920 patient records and 16 clinical attributes. After preprocessing and encoding categorical variables, the dataset was divided into training and testing sets using an 80–20 stratified split, resulting in 736 samples for training and 184 samples for testing.

Three machine learning algorithms were evaluated for heart disease prediction:

- Logistic Regression
- Random Forest
- Extreme Gradient Boosting (XGBoost)

Model performance was evaluated using accuracy, ROC-AUC score, and cross-validation metrics. These evaluation measures provide a comprehensive assessment of predictive capability and model generalization.

4.2 Model Performance Comparison

The performance of the implemented machine learning models is summarized in Table 2.

Model	Accuracy	ROC-AUC	Cross Validation
Logistic Regression	0.821	0.922	0.79

Random Forest	0.842	0.935	0.761
XGBoost	0.853	0.909	0.709

Table 2: Model Performance Comparison

The results indicate that XGBoost achieved the highest prediction accuracy of approximately 85.3%, outperforming both Logistic Regression and Random Forest models. Logistic Regression achieved an accuracy of 82.1%, while Random Forest achieved 84.2% accuracy.

Although Random Forest produced a slightly higher ROC-AUC score (0.935), the XGBoost model demonstrated better classification performance overall and was therefore selected as the best-performing model for further interpretability analysis.

These results are consistent with previous studies that highlight the effectiveness of gradient boosting algorithms for structured healthcare datasets.

4.3 Classification Performance

The classification performance of the selected XGBoost model was further evaluated using precision, recall, and F1-score metrics. The classification report is presented in Table 3.

Class	Precision	Recall	F1-score	Support
No Heart Disease (0)	0.87	0.79	0.83	82
Heart Disease (1)	0.84	0.90	0.87	102

Table 3: Classification Report for XGBoost Model

The model demonstrates strong performance in identifying patients with heart disease, achieving a recall of 0.90 for the positive class, indicating that the model correctly identifies a large proportion of patients who have cardiovascular disease. The overall F1-score of approximately 0.85 reflects a balanced trade-off between precision and recall.

These results indicate that the proposed predictive model is capable of effectively distinguishing between patients with and without heart disease.

4.4 Global Feature Importance Analysis

To analyze the overall behavior of the predictive model, global feature importance was computed using the trained XGBoost classifier. The analysis revealed several clinical attributes that significantly influence the prediction of heart disease.

The top contributing features identified by the model include:

Feature	Importance Score
Chest Pain Type (Asymptomatic)	0.206
Exercise Induced Angina	0.086
Chest Pain Type (Atypical Angina)	0.054
Dataset Source (Switzerland)	0.051
Thalassemia (Normal)	0.045
Sex (Female)	0.043
ST Segment Slope (Flat)	0.039
Dataset Source (VA Long Beach)	0.038
Number of Major Vessels	0.036
Oldpeak	0.031

Table 4: Global Feature Importance

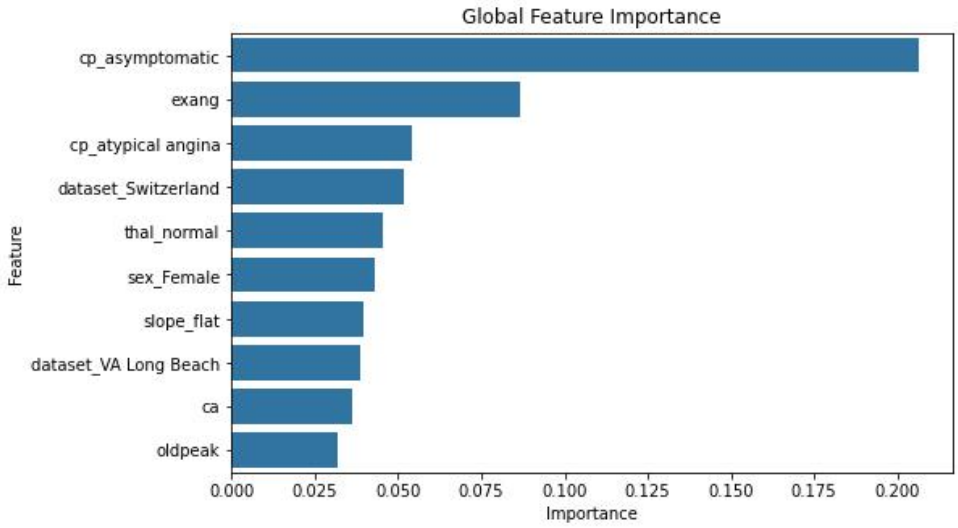

Figure 2: Global Feature Importance

The results highlight chest pain type, exercise-induced angina, and cardiac stress indicators as key predictors of cardiovascular disease. These attributes are widely recognized clinical indicators of heart disease, demonstrating that the model's predictions align with established medical knowledge.

Global feature importance analysis provides insights into which variables most strongly influence the model's predictions across the entire dataset.

4.5 Local Interpretability using LIME

While global feature importance provides an overall understanding of model behavior, it does not explain the reasoning behind individual predictions. To address this limitation, the LIME (Local Interpretable Model-Agnostic Explanations) technique was applied to generate explanations for specific predictions made by the XGBoost model.

For a sample patient prediction, LIME identified several influential features contributing to the classification outcome. The explanation indicated that the following features played a key role in the prediction:

- Dataset source (Switzerland)
- Chest pain type (asymptomatic)
- Number of major vessels (ca)
- Gender (female indicator)
- Cholesterol levels
- Chest pain type (atypical angina)
- Exercise-induced angina
- ST depression (oldpeak)

These features contributed to the model predicting the presence of heart disease with a probability of approximately 0.583 for the selected patient instance.

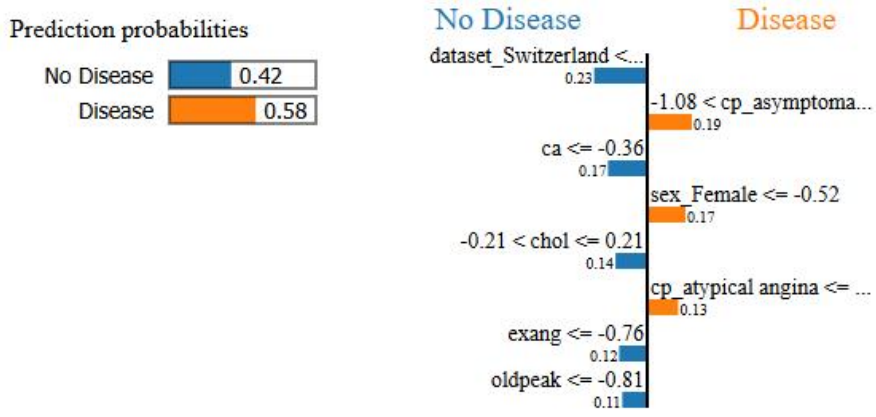

Figure 3: LIME explanations

LIME explanations provide a transparent view of the decision-making process by identifying which clinical attributes influenced the prediction outcome. Such explanations are particularly valuable in healthcare settings, where clinicians require patient-specific insights to validate AI-generated predictions.

4.6 Comparison of Global and Local Interpretability

An important aspect of explainable AI understands both global model behavior and local prediction reasoning. The comparison between global feature importance and LIME explanations revealed several overlapping attributes that influence predictions.

Global model analysis highlighted features such as:

- Chest pain type (asymptomatic)
- Exercise-induced angina
- Chest pain type (atypical angina)
- ST segment slope

Similarly, LIME explanations for individual predictions also emphasized features related to chest pain type, exercise-induced angina, and cardiovascular stress indicators.

The consistency between global and local explanations suggests that the model is capturing meaningful relationships within the dataset. This alignment between predictive modeling and clinical knowledge improves confidence in the reliability of the AI-based diagnostic system.

4.7 Discussion

The experimental results demonstrate that machine learning models can effectively predict heart disease using structured clinical datasets. Among the evaluated algorithms, XGBoost achieved the highest predictive accuracy, confirming its suitability for healthcare prediction tasks involving structured data.

However, predictive accuracy alone is insufficient for real-world medical applications. Clinicians require transparent explanations that justify the model's decisions and provide insights into the factors influencing predictions. By integrating LIME into the prediction framework, this study enhances model interpretability and enables clinicians to examine the reasoning behind individual predictions.

The combination of global feature importance and local LIME explanations provides a comprehensive understanding of model behavior. Global analysis reveals which features generally influence predictions, while LIME provides patient-specific explanations that support clinical decision-making.

Overall, the results demonstrate that combining machine learning prediction models with explainable AI techniques can significantly improve transparency and trust in AI-based healthcare systems.

5. CONCLUSION AND FUTURE WORK

This study presented an explainable machine learning framework for predicting heart disease using clinical data from the UCI Heart Disease dataset. The proposed approach integrates multiple machine learning algorithms with explainable artificial intelligence techniques to provide both accurate predictions and interpretable insights into the model's decision-making process.

Three classification algorithms: Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) were implemented and evaluated to determine the most effective predictive model. The experimental results demonstrated that the XGBoost model achieved the highest predictive accuracy of approximately 85.3%, outperforming the other models in overall classification performance. Additional evaluation using ROC-AUC and cross-validation confirmed the robustness and reliability of the predictive model.

Beyond predictive accuracy, the study focused on improving transparency and interpretability through the application of explainable AI techniques. Global feature importance analysis identified key clinical attributes influencing heart disease prediction, including chest pain type, exercise-induced angina, and ST-segment related indicators. These features correspond to medically relevant indicators of cardiovascular conditions, demonstrating that the machine learning model captures clinically meaningful relationships within the dataset.

To further enhance interpretability, the Local Interpretable Model-Agnostic Explanations (LIME) technique was used to generate patient-level explanations for individual predictions. LIME provided clear insights into how specific features influenced the model's decision for a particular patient instance. The comparison between global feature importance and local LIME explanations revealed a strong alignment between overall model behavior and individual prediction reasoning, thereby increasing confidence in the reliability of the predictive system.

The findings of this study demonstrate that integrating machine learning models with explainable AI techniques can significantly improve the transparency and trustworthiness of AI-based healthcare systems. Such systems have the potential to assist clinicians in diagnosing cardiovascular diseases and supporting data-driven clinical decision-making.

5.2 Future Work

Although the proposed framework demonstrates promising results for heart disease prediction and interpretability, several opportunities exist for further improvement and expansion.

Future research could explore the application of additional explainability techniques such as SHAP (SHapley Additive Explanations), ELIME, DLIME and counterfactual explanations to

provide deeper insights into model behavior. Comparing multiple explainability methods may help determine the most suitable approach for medical decision support systems.

Another direction for future work involves incorporating larger and more diverse healthcare datasets to improve the generalization capability of predictive models. Real-world clinical datasets with additional patient attributes could further enhance the reliability and clinical relevance of the system.

Furthermore, deep learning models and hybrid ensemble approaches may be investigated to determine whether they can achieve improved predictive performance while maintaining interpretability through advanced explainable AI frameworks.

Finally, integrating explainable machine learning models into clinical decision support systems could provide healthcare professionals with practical tools for risk assessment and early detection of cardiovascular diseases. Such systems could assist clinicians in interpreting patient data more effectively while maintaining transparency and accountability in medical decision-making.

References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
- Dua, D., & Graff, C. (2019). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- 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. https://doi.org/10.1145/2939672.2939778
- Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.
- Agarwal, D., Logeswari, P. (2025). Explainable AI in Cancer Diagnosis: Enhancing Interpretability with SHAP on Benign and Malignant Tumor Detection. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2025.66580
- Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z