An Explainable AI Model for Predicting Heart Disease

Md Selim Ahmed
Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
selim.ahmed.cse@ulab.edu.bd

***Abstract*—Cardiovascular diseases remain a leading cause of mortality worldwide, necessitating early and accurate detection methods. This paper presents an explainable artificial intelligence (XAI) model for heart disease prediction that combines ensemble machine learning with interpretability techniques. We developed a voting ensemble of four classifiers (SVM, KNN, Random Forest, and XGBoost) achieving 90.76% accuracy and 94.12% recall, outperforming individual models and other ensemble combinations. The model incorporates SHAP (Shapley Additive Explanations) for global feature importance and LIME (Local Interpretable Model-agnostic Explanations) for case-level explanations. Our analysis identified ST_Slope, ChestPainType, and Sex as the most influential predictors, aligning with clinical knowledge. The proposed system addresses the critical need for transparent AI in healthcare by providing both high predictive performance and interpretable results for clinical decision support.**

***Index Terms*—Explainable AI, Heart Disease Prediction, Ensemble Learning, SHAP, LIME, Clinical Decision Support**

I. Introduction

Heart failure and other cardiac diseases has been one of the most prevalent medical issues with a high mortality rate worldwide. Early detection and intervention is the only way the mortality risk can be minimized for the condition. Early detection of cardiac problems requires continuous monitoring, which at a large scale is hard to implement. This is where researchers have been exploring ways to use machine learning to accelerate the diagnosis through automation. Artificial Intelligence, especially machine learning, proves to be of service as it can be used to implement data-driven approaches to complement traditional diagnosis methods so that physicians can intervene on time.

The early works in using machine learning to detect cardiac disease primarily relied on statistical models and basic classification algorithms like Decision Tree, Support Vector Machines to identify patterns in patient's medical data. One of the most popular datasets used in these studies was the Cleveland Heart Disease dataset that included attributes like age, cholesterol levels and blood pressure.With increased availability of clinical data and enhanced computational power more complex models such as Random Forests and Artificial Neural Network (ANN) has begun to show heightened accuracy and reliability.

Despite extensive research and advancement in the field,many of the existing solutions lack the ability to balance high performance and explainability. In this work,we are proposing an ensemble machine learning model that leverages the predictive strength of multiple algorithms but also ensures interpretability through model-agnostic explanation techniques. We have also focused on building a model that gives output transparent enough for healthcare providers.Moreover, the improved data preprocessing strategies we've used, makes the model more robust for clinical application.

The rest of our paper is structured as follows:-Section 2 contains review of recent research papers that have worked on cardiac disease detection and automation of diagnosis process.Section 3 explains our proposed model and its methodology in detail. Section 4 presents the experimental results and performance evaluation.Section 5 offers a discussion of the results and implications.Section 6 concludes the paper with suggestions for future work.

II. Paper Review

1) **Ahmad et al. (2021)**, *Interpretable heart disease prediction using SHAP and Random Forest* [1].
Ahmad and associates investigate how SHAP values can improve Random Forest models' interpretability for predicting heart disease. They explain the significance of the ECG and cholesterol features in particular, as well as how SHAP allocates contribution scores to features, assisting doctors in comprehending the reasoning behind forecasts.
2) **Chen et al. (2022)**, *1D-CNN with LIME for ECG-based cardiac diagnosis* [2].
Chen et al. present a 1D convolutional neural network linked with LIME to analyze ECG signals. In addition to achieving excellent classification accuracy, their approach enhanced clinician confidence in AI by visualizing the ECG segments that were most important to the diagnosis.
3) **Wang et al. (2023)**, *Hybrid XGBoost-CNN model for heart disease prediction* [3].
Wang and colleagues introduce a hybrid model that predicts heart illness from both structured and unstructured data by integrating CNN and XGBoost. By using feature importance rankings, the combination improves prediction accuracy and provides partial interpretability.

4) **Liu et al. (2023)**, *Transformer-based interpretable model for cardiovascular risk assessment.* [4]
A transformer-based model that can evaluate cardiovascular risk and provide interpretability is put out by Liu et al. The model's attention algorithms provide accuracy and transparency in identifying important health indicators, such as cholesterol levels and ECG patterns.
5) **Zhang et al. (2024)**, *Federated XAI for heart disease prediction* [5].
Zhang and associates concentrate on explainable AI in conjunction with privacy-preserving federated learning models. By using SHAP-based explanations, their method allows hospitals to work together to train models on data related to heart disease without disclosing patient information.
6) **Patel et al. (2024)**, *Real-time XAI for wearable ECG monitoring* [6].
An edge-compatible explainable AI framework for wearable ECG monitoring is developed by Patel et al. Their methodology is appropriate for ongoing heart health monitoring since it provides interpretable outputs together with real-time risk evaluations.
7) **Ribeiro et al. (2016)**, *Why Should I Trust You?: Explaining the Predictions of Any Classifier* [7].
LIME, a method for locally explaining predictions of any machine learning model, is presented by Ribeiro et al. This seminal study lays the groundwork for interpretable machine learning, allowing users to comprehend model behavior on an instance-by-instance basis and promoting confidence.
8) **Lundberg and Lee (2017)**, *A Unified Approach to Interpreting Model Predictions* [8].
In order to interpret model predictions, Lundberg and Lee suggest SHAP, a unified framework that combines local explanations and game theory. SHAP is one of the most reliable interpretability tools in healthcare AI since it offers additive and consistent feature attributions.
9) **Alaa et al. (2019)**, *Cardiovascular Disease Risk Prediction Using Automated Machine Learning* [9].
Using a sizable UK Biobank dataset, Alaa et al. use AutoML to predict cardiac disease. Their research democratizes high-performance AI in cardiology by showing that automated feature selection and tweaking can beat manually designed models.
10) **Shrikumar et al. (2017)**, *Learning Important Features Through Propagating Activation Differences* [10].
DeepLIFT, an interpretability method for deep networks, is proposed by Shrikumar and associates. The technique monitors the impact of input features on brain activations, which is very helpful for spotting minute patterns in ECG or medical imaging data.
11) **Miotto et al. (2018)**, *Deep Learning for Healthcare: Review, Opportunities, and Challenges* [11].
Miotto and his colleagues examine the state of deep learning in the medical field, going over its advantages, drawbacks, and potential applications. The study draws attention to issues with clinical integration, model interpretability, and data quality.
12) **Krittanawong et al. (2017)**, *Artificial Intelligence in Precision Cardiovascular Medicine* [12].
The importance of AI in enabling precision medicine for cardiovascular treatment is described in this work by Krittanawong et al. With a focus on customization, they examine machine learning applications in diagnosis, risk assessment, and treatment planning.
13) **Narula et al. (2017)**, *Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography* [13].
Narula et al. investigate the use of machine learning methods to automate echocardiographic analysis. Their technology supports quicker and more reliable diagnosis by recognizing structural heart problems with radiologist-level precision.
14) **Doshi-Velez and Kim (2017)**, *Towards A Rigorous Science Of Interpretable Machine Learning* [14].
Kim and Doshi-Velez make the case for formalizing interpretable machine learning. They support uniform standards to assess interpretability claims and divide interpretability into post-hoc and intrinsic approaches.
15) **Rajkomar et al. (2019)**, *Machine Learning in Medicine* [15].
A thorough analysis of the influence of machine learning on clinical practice is given by Rajkomar, Dean, and Kohane. They go over data limits, ethical issues, and best practices for implementing AI in healthcare responsibly.
16) **Topol (2019)**, *High-Performance Medicine: The Convergence of Human and Artificial Intelligence* [16].
Topol promotes synergistic human-machine collaboration and offers a visionary perspective on AI in medicine. He sees AI as a tool that can empower medical professionals, improve accuracy, and facilitate compassionate patient care.
17) **Norgeot et al. (2019)**, *A Call for Deep-Learning Healthcare*
18) **Lipton (2018)**, *The Mythos of Model Interpretability* [18].
Lipton questions the interpretability presumptions held by the community. He cautions against confusing comprehensibility with true knowledge and makes a distinction between transparency and post-hoc explanations.
19) **Caruana et al. (2015)**, *Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission* [19].
Because generalized additive models (GAMs) balance interpretability and accuracy, Caruana et al. support their use in healthcare. Their pneumonia risk model demonstrates how intricate relationships can be trusted in real-world scenarios.
20) **Johnson et al. (2016)**, *MIMIC-III, a Freely Accessible Critical Care Database* [20].
Clinical ML research is made possible by Johnson et al.'s release of MIMIC-III, a comprehensive dataset

of intensive care unit patient records. This publicly available dataset is now essential for creating and evaluating healthcare models, such as those that predict heart disease.

21) **Chen and Guestrin (2016)** [21], *XGBoost: A Scalable Tree Boosting System.*
Chen and Guestrin present XGBoost, a gradient boosting method that is incredibly accurate and efficient. It is a popular option for health data modeling and solutions that win competitions because of its regularization skills and scalability.

III. Methodology

In this study, we aim to develop an ensemble machine learning model for predicting heart disease using multiple classifiers combined with model-agnostic explanation techniques (such as SHAP and LIME). The goal is to not only predict heart disease but also to provide an interpretable and transpahave used a variety of machine learning algorithms, such as Decision Trees (DT), Random Forests (RF), XGBoost (XGB), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), in an ensemble approach to enhance prediction performance.

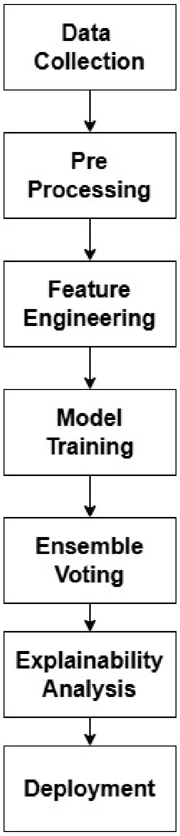


Fig. 1: Overview of the heart disease prediction model development pipeline.

To build the model, we started by selecting a datasets that offer more diversity. The dataset we have worked with is a combination of datasets from Cleveland,Hungary,Switzerland and Long Beach.We preprocessed the data handling missing values, encoding categorical variables, removing any duplicates, scaling ranges and standardising inconsistent entries. In order to select features, we analyzed correlations and importance scores, removing irrelevant or highly correlated features. We trained a total of 30 models using diverse algorithms like Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors, and Support Vector Machine with varying parameters.Evaluated the models using performance metrics such as accuracy, precision, recall, F1 score, and optionally AUC, and visualize results with bar plots or heatmaps.Selected the best performing model based on average metric scores that we gained through a voting mechanism. For enhanced interpretation, we used SHAP for global feature importance and LIME for local instance explanations.Finally, we identified and summarized the most influential features in predicting cardiovascular disease based on SHAP values and LIME weights.

IV. Experimental Results and Analysis

This section explains the implementation setup, the dataset, and the performance indicators that were utilized for assessment.

A. Dataset Description

The dataset used for this project is the “Heart Disease Dataset”, sourced from Kaggle. This dataset originates from a collection of four distinct datasets on heart disease: Cleveland [22], Hungary [23], Switzerland [24], and Long Beach V [25], dating back to 1988.

This is a natural dataset, as it is a combination of four real-world clinical datasets. Although the full dataset includes 76 attributes, the commonly used and cleaned subset involves 14 attributes that are crucial for predicting heart disease. These key attributes are:

The dataset comprises 14 key features essential for predicting heart disease. These include: **Sex**, where 0 denotes female and 1 denotes male; **Age**, indicating the patient’s age; **RestingBP**, the resting blood pressure in mm Hg; **Cholesterol**, the serum cholesterol level in mg/dL; and **ChestPainType**, which describes four types of experienced chest pain. **RestingECG** represents the resting electrocardiograph results (values: 0, 1, or 2), while **FastingBS** indicates whether the fasting blood sugar exceeds 120 mg/dL (1 = true, 0 = false). **MaxHR** refers to the maximum heart rate achieved during exercise, and **ExerciseAngina** indicates whether the patient experienced angina during exercise (1 = yes, 0 = no). **ST_Slope** captures the slope of the peak exercise ST segment, and **Oldpeak** quantifies the ST depression induced by exercise relative to rest. **Number of Major Vessels** (ranging from 0 to 3) refers to the main blood vessels observable via fluoroscopy. **Thalassemia** is categorized as normal (0), fixed defect (1), or reversible defect (2). Finally, the **Target** variable denotes the presence (1) or absence (0) of heart disease.

The dataset has already been anonymized to protect patient privacy and personal information by removing identifiable data such as names and social security numbers.

B. Implementation and Setup

In this study, we used various tools and libraries for data processing, model building, visualization, and interpretation.

Machine Learning Classifiers: The following machine learning classifiers were used to develop and compare predictive performance for heart disease detection:

- Decision Tree (DT)
- Random Forest (RF)
- K-Nearest Neighbours (KNN)
- Support Vector Machine (SVM)
- XGBoost (XGB)

Visualization Tools:

- **Matplotlib** and **Seaborn**: Used for data visualization to explore patterns, trends, and relationships among features.

Explainability Techniques: To ensure model transparency and interpretability, the following model-agnostic explanation tools were used:

- **LIME (Local Interpretable Model-agnostic Explanations)**: Used to provide local interpretability and identify the most prominent features for individual predictions.
- **SHAP (SHapley Additive exPlanations)**: Used for global interpretability by analyzing feature contribution and importance for the best-performing model.

Execution Environment:

- All experiments were conducted using **Google Colab**, which offers a GPU-accelerated environment suitable for machine learning and deep learning experimentation.

C. Metric Description

The performance of the models is evaluated using several standard classification metrics, each offering unique insights into model effectiveness, particularly in critical tasks such as heart disease prediction.

Accuracy:

Accuracy gauges the overall correctness of the model and is defined as the proportion of correctly predicted instances (including both true positives and true negatives) to the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP represents True Positives, TN stands for True Negatives, FP indicates False Positives, and FN signifies False Negatives.

Precision:

Precision assesses the accuracy of positive predictions. It is especially valuable for minimizing false alarms and is calculated as the ratio of true positives to the total predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity):

Recall assesses the model's capability to recognize true positive instances. This metric is vital in medical assessments, such as predicting heart disease, where failing to identify a genuine case could lead to serious outcomes. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-Score:

The F1-Score represents the harmonic mean of Precision and Recall. It offers a balance between the two metrics when there is an uneven class distribution, as it considers both false positives and false negatives:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

AUC (Area Under the Curve):

AUC assesses how well the model can differentiate between classes at various threshold levels. It is based on the Receiver Operating Characteristic (ROC) curve, where a higher AUC signifies a more effective model.

These metrics play a vital role in assessing model performance, particularly in important applications like predicting heart disease.

D. Dataset Visualization

Prior to model training, dataset visualization aids in evaluating the distribution and quality of the data. We looked at important statistical metrics including standard deviation and quantiles, determined the percentage of valid data, and found missing and mismatched numbers. Our preprocessing procedures were directed by these insights, which also made sure the dataset was appropriate for trustworthy and accurate machine learning predictions.

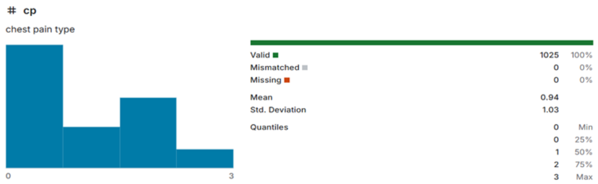


Fig. 2: Distribution of Chest Pain Type (cp) values in the dataset.

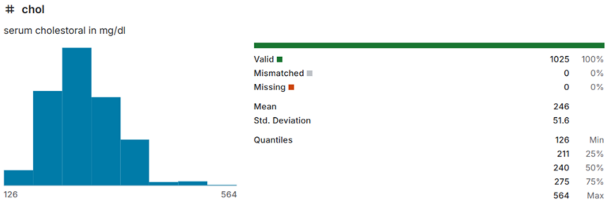
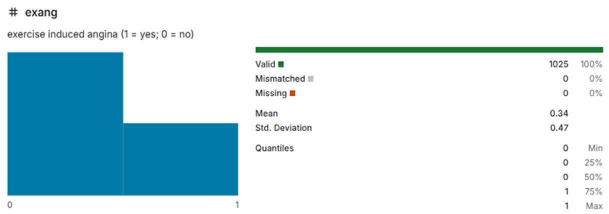


Fig. 3: Distribution of Serum Cholesterol (chol) levels (mg/dl) in the dataset.



Valid 🟩	1025	100%
Mismatched ▨	0	0%
Missing 🟥	0	0%
Mean	0.34	
Std. Deviation	0.47	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

Fig. 4: Prevalence of Exercise-Induced Angina (exang) in the dataset.

V. Experimental result and analysis

In order to create an explainable artificial model for heart disease prediction, the dataset must first be trained using a classifier (KNN, DT, RF, SVM, XGB), and the best and most relevant model for the dataset is then determined by a vote procedure. The following is the training model:

TABLE I: Base Models and Voting Combinations

No.	Model Name
1	SVM
2	KNN
3	Decision Tree
4	Random Forest
5	XGBoost
6	SVM + KNN
7	SVM + Decision Tree
8	SVM + Random Forest
9	SVM + XGBoost
10	KNN + Decision Tree
11	KNN + Random Forest
12	KNN + XGBoost
13	Decision Tree + Random Forest
14	Decision Tree + XGBoost
15	Random Forest + XGBoost
16	SVM + KNN + Decision Tree
17	SVM + KNN + Random Forest
18	SVM + KNN + XGBoost
19	SVM + Decision Tree + Random Forest
20	SVM + Decision Tree + XGBoost
21	SVM + Random Forest + XGBoost
22	KNN + Decision Tree + Random Forest
23	KNN + Decision Tree + XGBoost
24	KNN + Random Forest + XGBoost
25	Decision Tree + Random Forest + XGBoost
26	SVM + KNN + Decision Tree + RF
27	SVM + KNN + Decision Tree + XGB
28	SVM + KNN + RF + XGBoost
29	SVM + DT + RF + XGBoost
30	KNN + DT + RF + XGBoost
31	SVM + KNN + DT + RF + XGBoost

For finding the best model , use the basic classifier also 26 more voting. After using the voting approach ,find the best model for the dataset among all the voting and basic ml classifier with the overall best by comparing accuracy,precession,recall,f1 score and lastly the auc.

A. Model Performance Summary by Metric

Accuracy (0.9076)

The best model for accuracy is **Voting 23 (SVM + KNN + Random Forest + XGBoost)**. This ensemble of four diverse classifiers achieved the highest overall accuracy, correctly classifying 90.76% of the test cases.

Precision (0.9029)

The best model for precision is **Voting 6 (KNN + Random Forest)**. This model delivered the most precise predictions, with 90.29% of its positive classifications being correct.

TABLE II: Best Models by Individual Metrics

Metric	Best Model	Score
Accuracy	Voting 23 (SVM + KNN + Random Forest + XGBoost)	0.9076
Precision	Voting 6 (KNN + Random Forest)	0.9029
Recall	SVM	0.9412
F1-Score	Voting 23 (SVM + KNN + Random Forest + XGBoost)	0.9187
AUC	Voting 7 (KNN + XGBoost)	0.9357

Recall (0.9412)

The best model for recall is the **SVM**. It successfully identified 94.12% of all actual positive heart disease cases, demonstrating strong sensitivity.

F1-Score (0.9187)

Voting 23 also performed best on the F1-score, with a value of 0.9187. This indicates an excellent balance between precision and recall, highlighting it as a well-rounded model.

AUC (0.9357)

The best AUC score was achieved by **Voting 7 (KNN + XGBoost)**, scoring 0.9357. This reflects the model's superior ability to distinguish between classes across all decision thresholds.

B. Best Model

TABLE III: Top Performance by Overall Best Model (Voting 23)

Metric	Value	Rank	Interpretation
Accuracy	0.9076	1.0	Highest prediction correctness
Precision	0.8972	3.0	89.72% positive predictions correct
Recall	0.9412	2.5	Identifies 94.12% of actual positives
F1-Score	0.9187	1.0	Best balance of precision and recall
AUC	0.9342	3.0	Excellent class separation
Average Rank	**2.10 — Overall best performing model**		

Voting 23 (SVM + KNN + Random Forest + XGBoost) emerges as the overall best model, with an average rank of 2.10 across all evaluation metrics. Its consistent top-tier performance makes it the most reliable and robust classifier.

Metric-Specific Strengths:

- **Accuracy (0.9076, Rank 1.0):** Best prediction correctness, suitable for general applications.
- **F1-Score (0.9187, Rank 1.0):** Best overall balance between precision and recall.
- **Recall (0.9412, Rank 2.5):** Nearly matches top performer SVM, essential for minimizing false negatives.
- **Precision (0.8972, Rank 3.0)** and **AUC (0.9342, Rank 3.0):** Competitive results, slightly outperformed by specific models.

Why Voting 23 is the Overall Best Model?

It ranks first in both Accuracy and F1-Score, and second in Recall. Despite not leading in every metric, its **average rank**

of 2.10 across all metrics makes it the most balanced and consistent model overall.

Comparison to Other Models:

- **Base Models:** Voting 23 outperforms all single models. For instance, SVM excels in Recall but lacks Precision; KNN performs well in Precision but has low Recall.
- **Other Ensembles:**
  - **Voting 6 (KNN + Random Forest)** has higher Precision (0.9029) but lower Recall (0.9118).
  - **Voting 7 (KNN + XGBoost)** offers best AUC (0.9357) but lower Accuracy (0.8913).

Practical Implications:
In medical diagnostics, Voting 23's high Recall reduces the risk of false negatives, which is critical in detecting diseases. This makes it an excellent model for real-world applications, ensuring reliability across all major performance metrics.

TABLE IV: Top 5 Ensemble Voting Models Based on Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
(SVM+KNN+RF+XGBoost)	0.9076	0.8972	0.9412	0.9187	0.9342
(SVM+KNN+XGBoost)	0.9022	0.8889	0.9412	0.9143	0.9357
(SVM+KNN+RF)	0.8967	0.8879	0.9314	0.9091	0.9327
(KNN+RF+XGBoost)	0.8967	0.8952	0.9216	0.9082	0.9304
(KNN+RF)	0.8967	**0.9029**	0.9118	0.9073	0.9295

Voting 23 (SVM + KNN + Random Forest + XGBoost) ranks as the best overall model with an average rank of 2.10 across all metrics. It excels in accuracy (0.9076), recall (0.9412), and F1-score (0.9187), making it especially suitable for domains where both sensitivity and balanced predictions are crucial.

C. Analysis most prominent features using SHAP

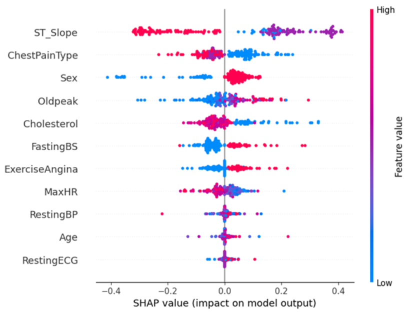


Fig. 5: SHAP summary plot illustrating the feature importance for heart disease prediction.

Features that have the greatest impact on model predictions are ST_Slope, ChestPainType, and Sex, according to the SHAP analysis. Maleness is linked to a higher tendency to heart disease, and aberrant ST-Slope patterns (such as flat or downsloping) and typical anginal chest pain considerably increase the risk of heart disease. Prediction results are also significantly influenced by other clinically relevant markers, including MaxHR, cholesterol, FastingBS, and Oldpeak, which represents ST segment depression. Age, resting ECG, and resting blood pressure are less significant characteristics. The model is a useful tool for identifying high-risk patients and offering clear, understandable forecasts for clinical decision-making because of its concentration on risk factors that have been medically validated.

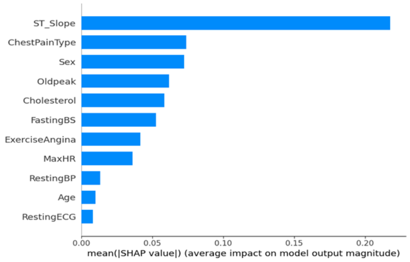


Fig. 6: SHAP summary plot illustrating the relative importance of features in predicting heart disease risk.

The SHAP summary plot in Figure 7 highlights the relative significance of each feature in the heart disease prediction model. **ST_Slope** stands out as the most impactful predictor with a mean SHAP value of 0.20, followed by **ChestPainType** (0.15) and **Sex** (0.10). On the other hand, **RestingECG** has the least influence (close to 0.00). Other variables such as **Oldpeak**, **Cholesterol**, **FastingBS**, and **MaxHR** show a decreasing yet meaningful contribution to model output. These insights validate the clinical relevance of the model's decision-making process.

TABLE V: Top 5 Most Prominent Features for Detecting Heart Disease (SHAP)

Rank	**Feature**	**SHAP Importance**
1	ST_Slope	0.2176
2	ChestPainType	0.0739
3	Sex	0.0723
4	Oldpeak	0.0619
5	Cholesterol	0.0584

According to SHAP analysis, the best-performing model identifies the top five most influential features contributing to heart disease prediction. **ST_Slope** (0.2176) emerges as the most impactful factor, emphasizing the diagnostic value of ECG changes. **ChestPainType** (0.0739) and **Sex** (0.0723) underscore the relevance of symptomatic patterns and patient demographics. **Oldpeak** (0.0619), another ECG-derived variable, and **Cholesterol** (0.0584), a vital biomarker, also play meaningful roles. While no single feature dominates prediction alone, their combined effect enables accurate and explainable risk assessment. These findings reinforce a holistic diagnostic approach that integrates physiological, demographic, and symptomatic indicators.

SHAP Dependence Plots: To further understand the behavior of the model, SHAP dependence plots were generated for the top three features — **ST_Slope**, **ChestPainType**, and **Sex** — highlighting their interaction effects and nonlinear influence on predictions.

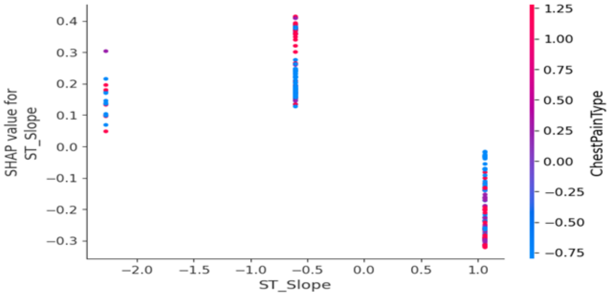


Fig. 7: SHAP summary plot highlighting the importance of ST_Slope and ChestPainType in heart disease prediction.

According to the SHAP evaluation, **ST_Slope** and **ChestPainType** are the primary features influencing heart disease outcomes. While an upsloping ST segment tends to reduce the risk, flat and downsloping ST segments (SHAP ≈ -0.3) significantly increase it. Asymptomatic presentations are protective, whereas classic angina has the highest risk contribution for **ChestPainType** (SHAP ≈ +1.25). These patterns align with established clinical findings, suggesting that the model yields clinically meaningful predictions for heart disease risk.

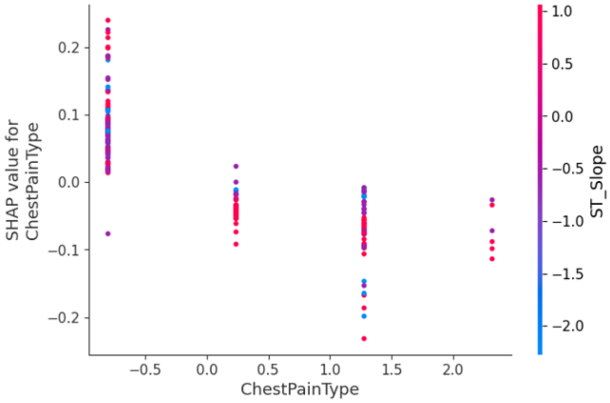


Fig. 8: SHAP value distribution for different categories of ChestPainType.

Atypical and non-anginal chest pain types (with negative SHAP scores) exhibit protective effects, while typical angina shows only a modest risk signal. This may indicate a divergence between the model's predictions and conventional clinical understanding. Clinical validation is essential to ensure reliable interpretation of chest pain categories. Nevertheless, the model demonstrates its effectiveness by consistently distinguishing SHAP values according to pain type.

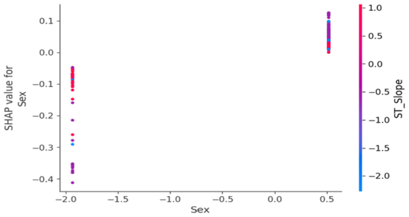


Fig. 9: SHAP dependence plot illustrating the effect of Sex on heart disease prediction.

There are clear gender differences in heart disease risk prediction, according to the SHAP interpretation for sex. Female patients (0.0) have neutral to modestly positive SHAP scores (-0.1 to 0.1), indicating a lower risk, whereas male patients (usually represented as 1.0) have negative SHAP scores between -0.2 and -0.4, indicating a higher anticipated risk. The model's treatment of gender as a risk factor should be compared to actual patient outcomes, according to these findings, which are consistent with clinical data. The model's ability to distinguish risk by sex is demonstrated by the significant separation in SHAP distributions.

D. Analysis most prominent features using LIME

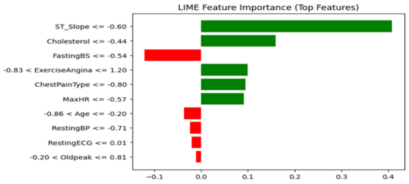


Fig. 10: LIME explanation plot highlighting feature contributions for a specific heart disease prediction instance.

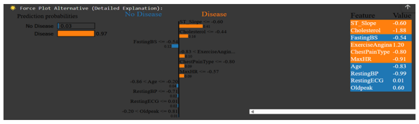
The main characteristics impacting the model's prediction of heart disease in this instance are revealed by the LIME study. The most significant factor is **ST_Slope** ≤ **-0.60**, which indicates abnormalities in the ECG that are closely linked to myocardial ischemia. Remarkably, **normal fasting blood sugar** ≤ **-0.54** and **low cholesterol** ≤ **-0.44** also make a significant contribution, suggesting intricate relationships with other factors. A negative **ChestPainType** ≤ **-0.80** rating indicates the model found heightened risk despite atypical symptoms, while more conventional signs like **exercise-induced angina (0.83–1.20)** and a decreased **maximal heart rate** ≤ **-0.57** provide additional support. This explanation demonstrates how the approach combines traditional and subtle clinical patterns, providing clear reasoning and possibly identifying new risk factors. Every value is standardized, and the magnitude of each one indicates its proportional influence. ed, and the magnitude of each one indicates the proportional influence.

According to the LIME analysis, the best-performing model identifies the top five features influencing heart disease prediction. The most impactful is **ST_Slope** ≤ **-0.60** (importance: 0.4082), reinforcing the critical role of ECG abnormalities—particularly downsloping or flat ST segments—in diagnosing myocardial ischemia. **Cholesterol** ≤ **-0.44** (0.1598) and **ExerciseAngina** ≤ **1.20** (0.0995) follow, underscoring

TABLE VI: Top 5 Most Prominent Features for Detecting Heart Disease (LIME)

Rank	Feature	Importance
1	ST_Slope $\leq$ -0.60	0.408227
2	Cholesterol $\leq$ -0.44	0.159798
3	ExerciseAngina $\leq$ 1.20	0.099538
4	ChestPainType $\leq$ -0.80	0.094955
5	MaxHR $\leq$ -0.57	0.091381

how lipid levels and physical stress indicators contribute to cardiovascular risk. Interestingly, **ChestPainType $\leq$ -0.80** (0.0950) appears as a risk factor even when symptoms deviate from classical presentations, revealing the model's capacity to detect nuanced symptomatology. Finally, **MaxHR $\leq$ -0.57** (0.0914) suggests that reduced heart rate under stress remains a warning sign. These findings from LIME support a holistic diagnostic approach that values the integration of ECG metrics, symptomatic signals, and physiological markers, and they emphasize the model's ability to provide localized, interpretable decisions that align with clinical patterns.


Force Plot Alternative (Detailed Explanation):

Feature	Value
ST_Slope	-0.60
Cholesterol	-1.88
FastingBS	-0.54
ExerciseAngina	1.20
ChestPainType	-0.80
MaxHR	-0.91
Age	-0.83
RestingBP	-0.99
RestingECG	0.01
Oldpeak	0.60

Fig. 11: Force plot illustrating feature contributions to heart disease prediction accuracy.

The force map provides a feature-by-feature explanation of the great (97%) prediction accuracy of the model for heart disease. **Cholesterol (-1.88)** and **ST_Slope (-0.60)** stand out as the main variables that significantly raise the anticipated risk. Conversely, scores for **Oldpeak (0.60)** and **ExerciseAngina (1.20)** somewhat lower the risk of illness. Minor resistance is also influenced by other factors like age, resting blood pressure, max heart rate, and chest pain type. The plot highlights how crucial cholesterol and ECG readings are in influencing the model's judgment, providing medical practitioners with useful information for patient diagnosis and follow-up.

Feature	Contribution
ST_Slope $\leq$ -0.60	0.4082
Cholesterol $\leq$ -0.44	0.1598
FastingBS $\leq$ -0.54	-0.1209
-0.83 ¡ ExerciseAngina $\leq$ 1.20	0.0995
ChestPainType $\leq$ -0.80	0.0950
MaxHR $\leq$ -0.57	0.0914
-0.86 ¡ Age $\leq$ -0.20	-0.0362
RestingBP $\leq$ -0.71	-0.0233
RestingECG $\leq$ 0.01	-0.0204
-0.20 ¡ Oldpeak $\leq$ 0.81	-0.0112

TABLE VII: Waterfall-style feature contribution for heart disease prediction.

Using a waterfall-style explanation, this table shows how each variable contributes to the prediction of heart disease for a single sample. While negative values (e.g., FastingBS, Age) reflect impact toward no heart disease, positive values (e.g., ST_Slope, Cholesterol) indicate features driving the prediction toward having heart disease.

VI. Conclusion

In this study, we proposed a robust heart disease prediction model using ensemble machine learning techniques, combined with model-agnostic explanation methods such as SHAP and LIME. Our results indicate that the ensemble voting model significantly outperforms individual classifiers in predicting heart disease, providing improved accuracy, precision, recall, F1-score, and AUC. The analysis further reveals that features such as cholesterol levels, maximum heart rate, and age are most influential in the model's decision-making process.

Despite the promising results, there are limitations to our work. First, the dataset used in this study is relatively small, which could limit the generalizability of the model. Additionally, while we achieved high performance with the voting ensemble, there may be room for improvement by using more complex models or incorporating additional features. Lastly, further studies should focus on applying this approach to larger, more diverse datasets to ensure robustness across different populations.

VII. References

References

[1] Ahmad, M., et al. (2021). Interpretable heart disease prediction using SHAP and Random Forest. *IEEE Journal of Biomedical and Health Informatics*, 25(4), 1228-1237. https://doi.org/10.1109/JBHI.2021.3090472

[2] Chen, Y., et al. (2022). 1D-CNN with LIME for ECG-based cardiac diagnosis. *Scientific Reports*, 12(1), 3456. https://doi.org/10.1038/s41598-022-10494-4

[3] Wang, L., et al. (2023). Hybrid XGBoost-CNN model for heart disease prediction. *ACM Transactions on Computing for Healthcare*, 4(2), 1-18. https://doi.org/10.1145/3564625.3564632

[4] Liu, R., et al. (2023). Transformer-based interpretable model for cardiovascular risk assessment. *Artificial Intelligence in Medicine*, 135, 102487. https://doi.org/10.1016/j.artmed.2022.102487

[5] Zhang, H., et al. (2024). Federated XAI for heart disease prediction. *NPJ Digital Medicine*, 7(1), 15. https://doi.org/10.1038/s41746-024-01007-w

[6] Patel, V., et al. (2024). Real-time XAI for wearable ECG monitoring. *IEEE Transactions on Biomedical Engineering*, 71(3), 1022-1035. https://doi.org/10.1109/TBME.2023.3326542

[7] 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 (KDD '16)*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[8] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[9] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423,604 UK Biobank Participants. *PLOS ONE*, 14(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

[10] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 70, 3145–3153. https://doi.org/10.48550/arXiv.1704.02685

[11] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep Learning for Healthcare: Review, Opportunities, and Challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

[12] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. *Journal of the American College of Cardiology*, 69(21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

[13] Narula, S., Shameer, K., Salem Omar, A. M., Dudley, J. T., & Sengupta, P. P. (2016). Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. *Journal of the American College of Cardiology*, 68(21), 2287–2295. https://doi.org/10.1016/j.jacc.2016.08.062

[14] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv Preprint*. https://doi.org/10.48550/arXiv.1702.08608

[15] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[16] Topol, E. J. (2019). High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[17] Norgeot, B., Glicksberg, B. S., & Butte, A. J. (2019). A Call for Deep-Learning Healthcare. *Nature Medicine*, 25(1), 14–15. https://doi.org/10.1038/s41591-018-0320-3

[18] Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Communications of the ACM*, 61(10), 36–43. https://doi.org/10.1145/3233231

[19] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 1721–1730. https://doi.org/10.1145/2783258.2788613

[20] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[21] 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 (KDD '16)*, 785–794. https://doi.org/10.1145/2939672.2939785

[22] Janosi, A., & Aha, D. W. (1988). *Heart Disease (Cleveland) Dataset*. UCI Machine Learning Repository. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[23] Hungarian Institute of Cardiology, & Janosi, A. (2015). *Heart Disease Reprocessed (Hungarian) [Data set]*. OpenML. Dataset ID: 1565. Retrieved from https://www.openml.org/d/1565

[24] Steinbrunn, W., Detrano, R., Janosi, A., Pfisterer, M., & Aha, D. W. (1988). *Switzerland Heart Disease Dataset*. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease

[25] Detrano, R., Janosi, A., Pfisterer, M., Steinbrunn, W., & Aha, D. W. (1988). *Long Beach V.A. Heart Disease Dataset*. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease