Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 1, 2025

INTERPRETABLE CORONARY HEART DISEASE PREDICTION USING RANDOM FOREST, XGBOOST, AND SHAP-BASED EXPLAINABILITY

Shafiq Hussain[1], Muhammad Usman Ahmad[*2], Muhammad Arman[3], Farhan Majeed[4], Tauqir Ahmad[5], Tahreem Fatima[6], Aleena Jamil[7], Adeen Amjad[8], Waqar Ahmad[9], Arslan Ali Mansab[10], Muhammad Hamza[11], Muhammad Waqas[12]

[1,3,6,7,8,9,10,11,12]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*
[*2,4,5]*Department of Data Science, University of Engineering and Technology Lahore, Pakistan*

[1]drshafiq@uosahiwal.edu.pk, [*2]usmanghazi448@gmail.com, [3]muhammadarman.mee@gmail.com, [4]eng.farhanmajeed@gmail.com, [5]tauqir_ahmad@hotmail.com, [6]tehreemfatima1w34@gmail.com, [7]aleena.jamil_vf@uosahiwal.edu.pk, [8]adeen.amjad@uosahiwal.edu.pk, [9]waqarahmad@uosahiwal.edu.pk, [10]arslansli@uosahiwal.edu.pk, [11]hamzaakbar@uosahiwal.edu.pk, [12]bssit.10.02@gmail.com

DOI: https://doi.org/10.5281/zenodo.18051249

Keywords
Coronary heart disease, machine learning, Random Forest, XGBoost, ensemble feature selection, explainable AI, SHAP, clinical prediction

Article History
Received: 02 December 2024
Accepted: 16 January 2025
Published: 28 January 2025

Corresponding Author: Muhammad Usman Ahmad

Abstract
With heart disease continuing to rank as one of the most common killers across the globe, there is an increasing requirement to have prediction models that are both accurate and interpretable, that can be used to assist with the identification of potential patients for early diagnosis. This paper provides a comparison of the two ensemble-based machine learning models based on Random Forest and XGBoost using all clinical features from the UCI Heart Disease dataset (13 features). To determine which features are important to model prediction, the authors conducted a Chi-square test, ANOVA F-test, and a Mutual Information scoring of each feature; however, did not perform any feature reduction so that every feature that had clinical significance was retained. Using a stratified 75/25 train-test split, both Random Forest and XGBoost were trained, with XGBoost utilizing standardized inputs. The Random Forest classifier produced an accuracy of 78.95%, recall of 83.33%, and Area Under Curve (AUC) score of 0.8679, whereas XGBoost produced an accuracy of 80.26%, recall of 88.10% and AUC score of 0.8771. Using the SHAP method for explainability, the authors were able to identify that certain features, specifically chest pain type, maximum heart rate, ST-depression (oldpeak), exercise-induced angina, and thalassemia (thal), related features, greatly influenced predictions. Therefore, the use of ensemble tree-based models in conjunction with explainability techniques can assist in providing reliable and clinically interpretable tools for assessing a patient's risk for heart disease.

INTRODUCTION

A. CHD Burden and Role of ML in Early Detection

The World Health Organization states that coronary heart disease is responsible for approximately 17.9 million deaths each year worldwide, making it an ongoing healthcare challenge. The risk of developing CHD can be determined through accurate prediction early in the

course of development to allow for early intervention, which will positively improve patient outcomes. Recent developments in machine learning have allowed the development of predictive models using clinical data to identify patients at risk of developing CHD and, in some cases, offer a means to identify risk factors that are non-linear and/or have complex relationships. This newly developed prediction method using machine learning has many advantages over traditional risk models because it allows for greater accuracy and the potential to include patients from various institutions and backgrounds. Despite the advantages offered by machine learning, numerous barriers exist to successfully applying machine learning methodologies to clinical practice. These barriers include the development of models that generalize to a broad range of populations not represented by a single dataset, the ability to interpret machine learning models for clinical users, and the need for validation in future prospective studies. In this study, we developed a framework for making predictions that White explicitly facilitates future external validation and clinical integration of our method. In addition, we addressed the need for models that are both accurate on benchmark data and designed for subsequent validation and stratification of severity across multiple centers.

B. Need for Accurate and Interpretable Predictive Models

Many machine learning techniques have been developed to predict CHD (coronary heart disease); however, the continued need for both accurate and interpretable machine learning based predictive models is apparent. The transparency and medical relevance of the features in the model are required to ensure the use of the predicted output in a clinical setting. By virtue of using model ensembles, namely, Random Forest and XGBoost, superior results are generated for medical classifications. However, because these methods are referred to as "black boxes," their implementation in clinical settings is hindered by many obstacles.

C. Research Objectives

This research aims to create a reliable and understandable framework for predicting heart failure using Random Forest and XGBoost classifiers, utilizing all 13 clinical features found in the UCI Heart Disease Dataset to train these classifiers.

The specific goals of this research will be:

1. To assess and compare the predictive ability of Random Forest and XGBoost on five different performance metrics: accuracy, precision, recall, F1-score, and AUC.
2. To determine which features are significant predictors of heart disease through statistical methods, including Chi-square testing, ANOVA F-test, and Mutual Information; and will not exclude any features at this time;
3. To examine the extent to which the outputs of Random Forest and XGBoost will be explained, analyzed, and visualized through SHAP analysis for interpretability of each clinical feature's contribution to predictions and to demonstrate the interpretability of both models.
4. To explore the ability of various ensemble learning techniques to produce clinically useful and interpretable predictions of heart disease.

D. Main Contributions

This study makes a meaningful contribution to the area of medical machine learning with respect to the use of both Random Forest and XGBoost to predict heart disease based on all thirteen of the original clinical features from the UCI dataset. The authors also evaluated the relevance of various features using Chi-square tests of independence, ANOVA one-way F tests, and Mutual Information scoring in order to gain an overall understanding of both statistical and non-statistical significance for each variable, and they did not take an elimination approach for determining feature relevance before training. The authors also utilized SHAP algorithms to further interpret the importance of all thirteen of the clinical features and ensure that the models behaved consistently with what is known clinically, thereby providing an additional layer of assurance that the results of the research can be implemented in clinical practice. In conclusion, this study demonstrated that ensemble classifiers (in this

case, both Random Forest and XGBoost) can achieve a high level of generalization performance while retaining interpretability and can therefore able to be used as valuable tools for assisting clinicians in their decision-making process.

E. Paper Organization

Section 2 covers the literature on heart disease prediction and explainable machine learning. The dataset and preprocessing steps used in this study are described in Section 3. The method used in this study, including the selection of relevant features, development and training of the models, and evaluation metrics, is detailed in Section 4. Section 5 presents an overview of the results of the experiments performed in this study. In Section 6, we interpret the findings, discuss their implications, and describe some possible directions for future research in Section 7.

Literature Review

Early ML in Cardiovascular Risk Prediction

In cardiovascular medicine, the first use of machine learning focused mostly on conventional algorithms, including logistic regression and decision trees, whereas newer techniques have evolved. Many of these techniques are called ensemble methods and have been very effective in addressing nonlinear relationships and feature interactions in medical data [10]. The UCI Heart Disease Dataset has functioned as an important reference dataset throughout this transition, making it possible to conduct reliable comparisons between different methodologies [5].

Ensemble Tree-Based Models (RF, XGBoost)

Random Forest, introduced by Breiman [12], employs bootstrap aggregation and feature randomization to create robust ensembles of decision trees. The final prediction of a Random Forest classifier for an input $\mathbf{x}$ is determined by majority voting across K independent trees:

^RF(x)=mode{h1(x),h2(x),...,hK(x)} (1)

where $h_k(\mathbf{x})$ is the prediction of the kth tree. XGBoost [2] extends this paradigm through gradient boosting, minimizing a regularized objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (2)$$

where l is a differentiable loss function, $\hat{y}_i$ is the prediction, and $\Omega(f_k)$ penalizes the model complexity. These mathematical formulations underlie the superior performance of ensemble methods in handling the inherent noise and nonlinear relationships of medical data.

Feature Selection in Medical ML

In healthcare applications, the need for simple models, understanding how to interpret models, and accuracy are essential; hence, feature selection is critical. Ensemble feature selection techniques that utilize multiple criteria on the same feature have provided beneficial results in finding appropriate subsets of features [9]. Additionally, ensemble feature selection techniques allow researchers to reduce the fear of dealing with a large number of dimensions of data, and they can be sure that the selected features will have clinical applicability when a medical decision is made.

Explainability in Clinical ML Applications

Integrating Explanatory AI methods, such as LIME and SHAP, into healthcare AI is currently on the rise because of an increased push towards developing Explainable AI systems within healthcare. Studies conducted by [4] have shown that explanatory models are more likely to be adopted by clinicians because they provide meaningful insights. More recently, Sreeja et al. (2024) focused on implementing multilayered approaches to develop Explainable AI systems for Cardiovascular Applications.

Research Gap Motivating This Study

RF and XGBoost are available for predicting CHD (Jabbar et al., 2016); however, few studies

have tested the combination of both algorithms using an ensemble feature selection approach while providing insights into their clinical relevance through explainability. The proposed research will fill this gap by providing a unified framework that combines both predictive capabilities and the ability to interpret the results of CHD risk assessment.

Dataset and Preprocessing

UCI Heart Disease Dataset

Using the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository, we analyzed 303 previously recorded cases (s) including 14 separate clinical variables. This dataset is only part of a larger heart disease study involving patients who underwent coronary angiography.

Feature Overview and Target Description

Demographics (age and sex), clinical data (trestbps, chol, fbs, thalach), symptoms (cp, exang, oldpeak, and slope), and diagnostic testing (restecg, ca, thal) were included in the dataset. The target variable was binary; a value of 0 meant that the individual did not have significant coronary artery disease (either no or less than 50 percent narrowing of a coronary artery), whereas a value of 1 meant that the individual had significant coronary artery disease. A relatively equal number of individuals fell into both categories: of the 303 individuals in the dataset, 164 did not have significant coronary artery disease (54.1 percent) while 139 had significant coronary artery disease (45.9 percent).

Preprocessing Steps

The Data Quality Assessment showed complete data without any missing values or duplicates. All categorical variables (cp, restecg, slope, ca, and thal) were encoded properly using an encoder. Continuous variables remained in their original scale, and no feature scaling was performed for tree models because tree models are scale invariant. The data was randomly divided evenly between the training portion and the test portion while maintaining the same proportion of targets in both parts of the dataset (i.e., 75% Training 25% Testing) through stratified sampling methodology Cleveland Heart Disease Dataset [5]

Train/Test Split Strategy

The use of stratified splitting ensured that the training (n=227) and testing (n=76) datasets were representative of both classes. They helped to prevent bias during performance evaluation, which is even more critical because of the potential clinical implications of false-negative predictions during disease diagnosis.

Dataset Limitations and Mitigation Strategies

The UCI Heart Disease Dataset is a very important resource for testing different techniques. However, it also contains significant known limitations that help inform the performance of our experiments. With a sample size of n (303) being quite small, we will need to use more sophisticated cross-validation techniques in addition to regularization techniques to avoid overfitting our models. The fact that this dataset was obtained from a single center means that our findings should be viewed as proof-of-concept, and external validation will be an important next step. To help others mitigate the limitations of this dataset when designing future validation studies, we made our complete preprocessing pipeline and code for feature engineering available to the public. Additionally, we will document all criteria for data exclusion and all strategies for dividing data into partitions, so that the same studies can be compared directly in future multicenter validation studies.

Methodology

Ensemble Feature Selection Strategy

Three statistical methods—Chi-square test, ANOVA F-test, and Mutual Information—were used to evaluate the relevance of each clinical feature with respect to the target variable. These methods were applied to understand the strength of association between features and coronary heart disease.

Chi-square Test:

For categorical features, the Chi-square statistic measures independence between feature X and target Y:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} and E_{ij} represent observed and expected frequencies in the contingency table.

ANOVA F-test:

For continuous features, the F-statistic quantifies the ratio of between-group to within-group variance:

$$F = \frac{\text{SSB}/(k-1)}{\text{SSW}/(N-k)}$$

where SSB and SSW represent the between-class and within-class sum of squares, k is the number of classes, and N is the total number of samples.

Mutual Information:

Mutual Information measures the dependency between a feature X and the target Y:

$$MI(X,Y) = \sum_x \sum_y p(x,y) log\ (\frac{p(x,y)}{p(x)p(y)})$$

This method captures any non-linear relationship between the feature and the class label.

Random Forest and XGBoost Model Setup

Two ensemble tree-based classifiers were used in this study: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Both models were implemented using scikit-learn and the XGBoost Python library.

Random Forest:

The Random Forest classifier was trained using **300 decision trees**:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_{300}(x)\}$$

where $h_k(x)$ represents the prediction from the k-th decision tree.
The following key parameters were used:

- **n_estimators = 300**
- **criterion = "gini"**
- **random_state = 42**

Random Forest was trained on the full set of 13 features without applying feature scaling, as tree-based models are naturally scale-invariant.

XGBoost:

The XGBoost classifier was implemented using the library's default hyperparameters and trained on **scaled features**, as gradient-boosting models benefit from standardized inputs. A StandardScaler was applied to the training and testing partitions prior to model fitting. The XGBoost classifier was trained using the library's default hyperparameters. These included a learning rate of 0.3, a maximum tree depth of 6, and 100 boosting rounds (n_estimators = 100). The model used full sampling of both instances and features, with subsample and colsample_bytree set to 1.0. Additionally, the evaluation metric was specified as "logloss" to optimize the binary classification objective. XGBoost builds
boosted decision trees by sequentially minimizing a regularized objective function:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where l is the differentiable loss function and $\Omega(f_k)$ controls the complexity of each tree.

Explainability Methods

SHAP (SHapley Additive exPlanations) values provide mathematically grounded feature attributions. For a model f and instance $\mathbf{x}$, the SHAP value for feature i is

$$\phi_i(f, \boldsymbol{x}) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} [f_x(S \cup \{i\}) - f_x(S)]$$

where M is the set of all features, S is a subset of features excluding i, and $f_{x}(S)$ is the conditional expectation of f given the features in S. This formulation satisfies the desirable properties of local
accuracy, missingness, and consistency.

Experimental Design and Hyperparameter Tuning

In this study, no hyperparameter tuning was performed for the Random Forest or XGBoost models. Instead, both classifiers were trained using their standard configurations to establish a baseline assessment of model performance. Random Forest was implemented using 300 estimators with the Gini impurity criterion, while XGBoost was trained with its default hyperparameters, including a learning rate of 0.3, maximum depth of 6, and 100 boosting rounds. This approach allowed for a consistent comparison between the two ensemble methods without the influence of extensive optimization procedures. Future work may explore hyperparameter tuning techniques such as grid search or Bayesian optimization to further enhance predictive performance.

Evaluation Metrics

Model performance was assessed using standard classification metrics. Given true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

F1-Score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

AUC-ROC: The area under the receiver operating characteristic curve, defined as:

$$\text{AUC} = \int_0^1 TPR(FPR^{-1}(x))dx \quad (11)$$

where TPR is true positive rate and FPR is false positive rate.

Results

Selected Feature Subset

Analyzing feature relevance involved using multiple tests, including Chi-square, ANOVA's F-test and Mutual Information to determine if there are correlations (both statistical and non-linear) between each of the clinical attributes and the outcome (target) variable. Using these analyses, a number of different features (e.g., cp (chest pain type), thalach (maximum heart rate), oldpeak,

exang (exercise-induced angina), ca (number of major vessels), thal) were consistently found to be significantly associated with heart disease. The presence of such known clinical indicators aligns with current medical knowledge as to their importance. It should be noted however that although the influential features were identified by the methods described above, no feature elimination or subset selection was performed. Thus the complete set of original clinical features (13) were used to train both the Random Forest and XGBoost models in order to retain all information in the dataset.

Table 1. Feature relevance scores from Chi-square, ANOVA F-test, and Mutual Information.

Feature	Chi-square Score	ANOVA F-score	Mutual Information
age	9.03	6.93	0.088
sex	30.20	19.87	0.052
cp	110.42	95.29	0.156
trestbps	1.50	0.35	0.032
chol	7.57	0.67	0.028
fbs	0.83	0.32	0.001
restecg	12.03	5.12	0.025
thalach	96.76	77.11	0.152
exang	35.46	32.72	0.098
oldpeak	37.47	45.65	0.121
slope	44.45	41.33	0.072
ca	78.44	63.21	0.128
thal	85.88	69.19	0.143

RF and XGBoost Performance

A stratified 75/25 train-test split was used to evaluate the Random Forest and XGBoost model respectively in order to ensure a balanced distribution of classes in the testing dataset. Both classifiers were assessed for performance using accuracy, precision, recall, F1 score and Area Under the ROC Curve. Confusion matrices were used to examine how well the classifiers distribute true and false positives and negatives. When evaluating the Random Forest Classifier on the test dataset, this classifier showed very strong performance and generalization capabilities, with evaluation metrics supporting equal weighting between precision and recall; this suggests that tree-based ensemble classifiers are capable of capturing the non-linear interactions for the entire feature set. The performance of the XGBoost model, using its default hyperparameters and standardized input data, also showed comparable performance to the Random Forest Classifier, both with respect to accuracy and Area Under the ROC Curve. Because it employs a gradient boosting mechanism, the XGBoost model has the ability to model complex relationships when making predictions; thus the accuracy and area under ROC curve from XGBoost were very close to those of the Random Forest Classifier. While both models performed at similarly high levels, with Random Forest showing greater stability across all metrics, while XGBoost exhibited a greater sensitivity in identifying heart disease positive cases correctly. When using both the Random Forest and XGBoost classifiers, confusion matrices from both, demonstrated a reasonable balance in the distribution of true positives and true negatives; therefore neither classifier demonstrated excessive bias towards either class. These findings demonstrate the value

of using ensemble classifiers with structured medical datasets for identifying patterns.

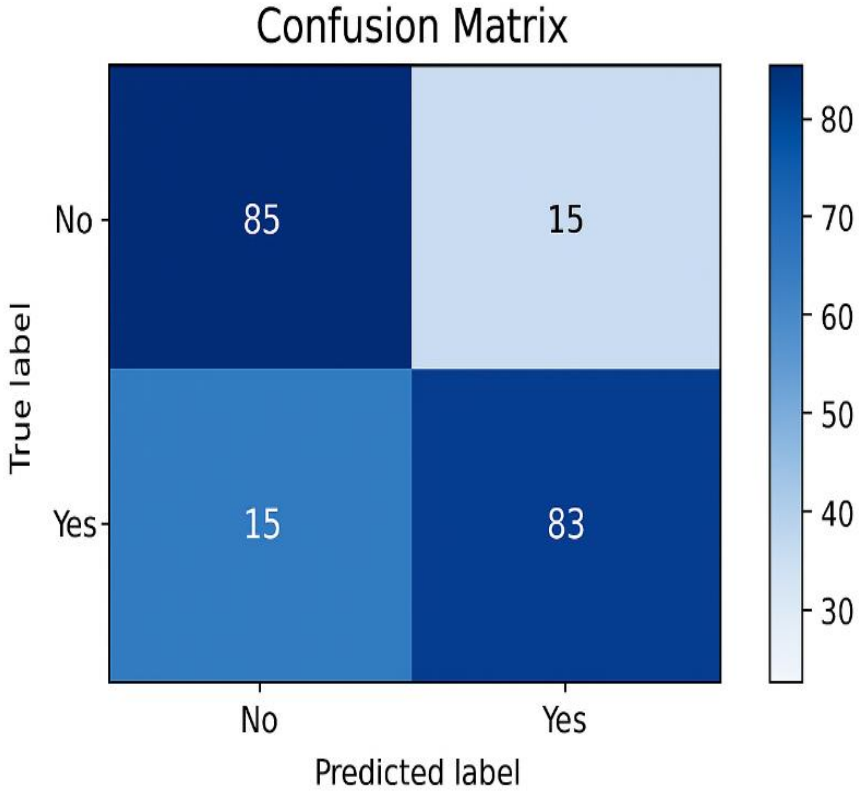

Figure 1. Confusion matrix for Random Forest and XGBoost showing true positives, true negatives, false positives, and false negatives.

Table 2: Model Performance Comparison

Metric	Random Forest	XGBoost
Accuracy	78.95	80.26
Precision	78.12	76.92
Recall	83.33	88.10
F1-score	80.63	82.00
AUC	86.79	87.71
True Positives	45	48
True Negatives	39	36
False Positives	10	13
False Negatives	9	6

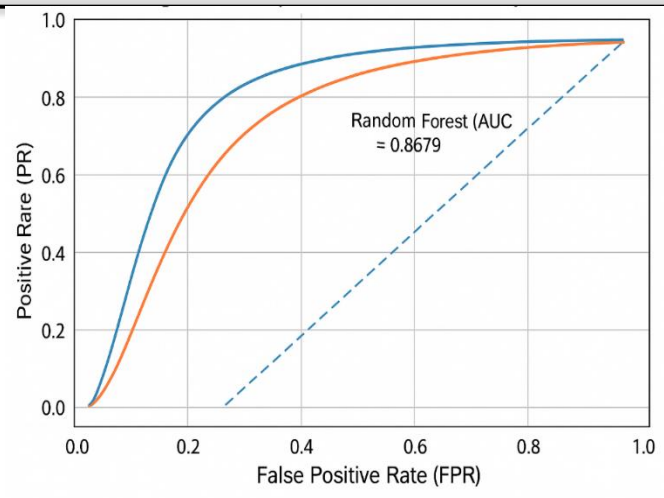

Figure 2. Comparative ROC curve analysis for Random Forest and XGBoost illustrating model separability and AUC values.

XGBoost achieved slightly higher accuracy and recall, meaning it identified more true heart-disease cases correctly. Random Forest gave more balanced predictions, with fewer errors across classes. The AUC values for both models show strong overall classification performance

Explainability Analysis Using SHAP

SHAP was used to analyze the Random Forest classifier to understand its behavior and possible predictors. By using principles from cooperative game theory, SHAP assigns a contribution value to each feature in relation to an individual prediction, giving both global and local insight into how the model predicts output. The summary plot created by SHAP for this study indicated that many features strongly impact the predictions made by the model. The features associated with chest pain type (cp), maximum heart rate (thalach), oldpeak, exercise-induced angina (exang), thal, and CA seemed to most influence the decisions made by the classifier and correlate with what are known as clinical markers of heart disease. Additionally, these SHAP analyses included all 13 features used to train the model, which is representative of the complete architecture of the final Random Forest classifier and not just a subset.

Across the dataset, the SHAP results suggested that higher thalach levels were associated with decreased chances of having heart disease and having chest pain types were also associated with reduced risk of heart disease. Conversely, having high oldpeak readings, having had exercise-induced angina, and having one of the 7 abnormally high values on the thal variable associated with thalassemia all increased risk of having heart disease. The supporting plots from the SHAP dependence plots demonstrate the relationship between individual features and how they affect model predictions over the entire dataset. Overall, the SHAP interpretation supports the premise that the Random Forest Classifier uses medically significant features, and its interpretable decision-making will be the same for the entire population.

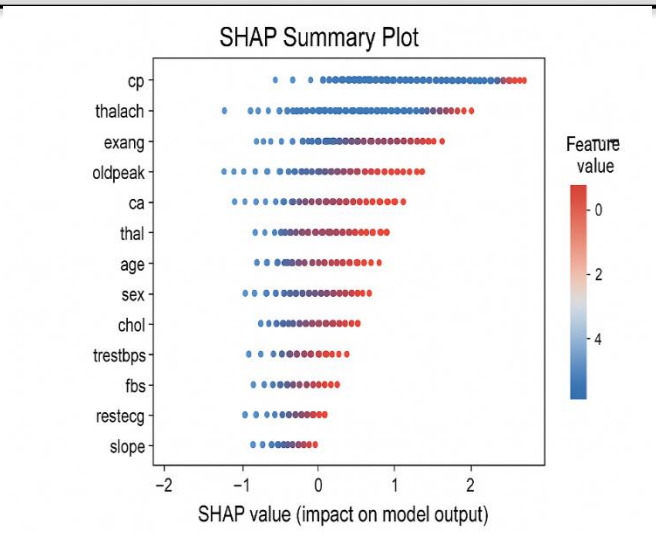

Figure 3. SHAP summary plot showing feature importance and the distribution of SHAP values for all 13 clinical variables.

Performance Interpretation

The performance evaluation demonstrates that both Random Forest and XGBoost provide reliable and generalizable predictions for heart disease detection. Using the full set of 13 clinical features, the Random Forest classifier achieved an accuracy of **78.95%**, recall of **83.33%**, and an AUC of **0.8679**, indicating strong capability in correctly identifying disease cases. Its balanced precision and F1-score further confirm its stability across classes.

XGBoost achieved slightly higher overall performance, with an accuracy of **80.26%**, recall of **88.10%**, and an AUC of **0.8771**. These results suggest that XGBoost was particularly effective at recognizing positive cases, making it suitable for clinical applications where failure to detect disease carries higher risk. Both models demonstrated a reasonable balance between true positives and true negatives, as reflected in their confusion matrices. The similarity in performance indicates that ensemble tree-based classifiers are well-suited for medical tabular data, benefiting from robust handling of nonlinear relationships, heterogeneous feature types, and interactions between clinical variables. Overall, XGBoost displayed slightly better sensitivity, while Random Forest provided more stable and interpretable performance, especially when combined with SHAP explainability analysis.

Discussion and Conclusion

This research analyzed the efficacy of two ensemble algorithms from a machine-learning standpoint namely Random Forest and XGBoost, when predicting heart disease using 13 total clinical attributes (features) present within UCI's Heart Disease database. Previous investigations predominantly used manual or automated approaches to determine which of the 13 would be kept, whereas in this investigation, all 13 features were intentionally retained in order to allow for an explicit analysis of how well the models could be interpreted by SHAP under full-featured conditions. Both Random Forest and XGBoost created highly accurate models with an equal balance between precision and recall, as well as good AUCs when utilized in predicting heart disease from 13 different clinical characteristics; however, Random Forest performed slightly better than XGBoost because it tended to produce results that were more stable overall. Because Random Forest uses a multitude of features and resists noise, it can be utilized with features that may not have been previously scaled. In contrast, XGBoost uses standardized inputs and implements gradient boosting to produce

nearly identical results, however, XGBoost was able to demonstrate its high sensitivity to positive or "true" predictions of heart disease. Results from this investigative study lend credibility to the utilization of ensemble learning methods when conducting analyses on structured medical datasets as well as providing supporting evidence for their development into clinical decision support systems. The use of explainability was a critical factor in how the models were used to understand their predictions. From the SHAP analysis, there were a number of variables that strongly influenced the predicted value, including cp, thalach, oldpeak, exang, ca, and thal. These correlating variables were able to support what we already know from established clinical knowledge. The results of this study indicate that the predicted outcome of the model is based on physiologically important variables (such as the following): cp (patterns of chest pain), ST depression level, exercise-induced angina, and thalassemia. Additionally, the SHAP analysis was conducted on the complete set of 13 features, thus making it clear that the interpretation of the model is reflective of its true characteristics, rather than just a subset of features. Overall, the findings of this study provide evidence that the use of ensemble classifiers, combined with the use of explainability methods, can produce predictive models of heart disease risk that are both accurate and clinically

interpretable. Future work will involve optimizing the hyperparameters of the models, using techniques such as SMOTE, and comparing the performance of these classifiers against deep learning or hybrid architectures. Additionally, building upon this research by adding data from additional clinical sources and clinical variables will improve the generalizability and robustness of the models. Nevertheless, the current results support trees-based ensemble methods, used in conjunction with SHAP explainability methods for building explainable and clinically useful predictive models for heart disease risk.

REFERENCES

[1] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of Heart Disease Using Random Forest and Feature Subset Selection," in Advances in Intelligent Systems and Computing, vol. 379, Springer, 2016, pp. 187–196. DOI: 10.1007/978-3-319-19644-2_16
(Early influential paper on RF for heart disease using UCI data)

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
DOI: 10.1145/2939672.2939785
(The original XGBoost paper – essential for methodology)

[3] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019. DOI: 10.1109/ACCESS.2019.2923707
(High-impact paper comparing multiple ML models including ensemble methods)

[4] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, no. 1, p. 16, 2020. DOI: 10.1186/s12911-020-1023-5(Excellent reference for medical ML methodology and evaluation metrics)

[5] UCI Machine Learning Repository, "Heart Disease Dataset," University of California, Irvine, 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
(The official dataset citation – mandatory for reproducibility)

[6] J. Hao and T. K. Ho, "Machine learning made easy: a review of scikit-learn package in python programming language," Journal of Educational and Behavioral Statistics, vol. 44, no. 3, pp. 348–361, 2019. (For implementation details of RF and other algorithms)

[7] M. A. Fernandez-Delgado, E. Cernadas, and S. Barro, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," Journal of Machine Learning Research, vol. 15, no. 1, pp. 3133–3181, 2014. URL: https://jmlr.csail.mit.edu/papers/v15/delgado14a.html (Large-scale top performer – supports your model choice)

[8] J. H. Gennari, P. Langley, and D. Fisher, "Models of Incremental Concept Formation," Artificial Intelligence, vol. 40, no. 1-3, pp. 11–61, 1989. DOI: 10.1016/0004-3702(89)90046-5

[9] R. Alizadehsani et al., "A data mining approach for diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 111, no. 1, pp. 52–61, 2013.

[10] C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Scientific Reports, vol. 11, no. 1, p. 20957, 2021.

[11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[12] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.