THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 5. Issue 2 (2025)
https://doi.org/10.62019/abbdm.v5i2.335

ASIAN BULLETIN OF BIG DATA MANAGEMENT
http://abbdm.com/
ISSN (Print): 2959-0795
ISSN (online): 2959-0809

An Explainable and Accurate Machine Learning Approach for Early Heart Disease Prediction Using Feature Selection and Ensemble Techniques

Khaliq ahmed*, Khalid bin Muhammad*, Malik zohaib Hussain, Abdul Khaliq*

Chronicle

Article history
Received: march 2, 2025
Received in the revised format: March 28, 2025
Accepted: April 28, 2025
Available online: May 30, 2025

Khaliq ahmed is currently affiliated with the Department of computer science Iqra university Karachi, Pakistan.
Email: Khaliq@iqra.edu.pk

Khalid bin Muhammad is currently affiliated with the faculty of engineering sciences and technology, department of computer science, Ziauddin university Karachi, Pakistan.
Email: Khalid.muhammad@zu.edu.pk

Malik zohaib Hussain & Abdul Khaliq are currently affiliated with the CCSIS, Institute of Business Management, Karachi, Pakistan.
Email: Zohaib.hussain@iobm.edu.pk
Email: khaliq@iobm.edu.pk

Corresponding Author*

Abstract

This research introduces a holistic machine learning-based approach to early heart disease prediction, utilizing state-of-the-art ensemble methods and explainable artificial intelligence (XAI). The envisioned model pipeline includes feature selection processes to improve prediction performance and interpretability. Ensemble techniques outperformed more classical models such as Logistic Regression and SVM, with different classifiers such as Random Forest, Gradient Boosting, and XGBoost being thoroughly compared. The best accuracy of 98.54% was attained by Random Forest, Gradient Boosting, and XGBoost, showing the effectiveness of ensemble methods in working with healthcare datasets. The precision and recall measures also drifted close to 1.0, indicating very few false negatives and false positives—essential for medical diagnoses. The AUC measures also supported the strength of the classifiers, with Random Forest showing a perfect 1.0. Visual outcomes confirm the adherence and performance of the suggested methodology in the most significant key performance indicators. This work prioritizes not just predictive performance but also explainability, so the model's choice can be understood by doctors. By combining accuracy with interpretability, this framework offers an accurate decision support system for cardiologists that allows for the early diagnosis and customized treatment plan. The visual analytics offered in the proposed work section also further support the practical relevance and clinical promise of the introduced method.

Keywords: heart diseases, machine leaning, Random Forest, XGBoost.

INTRODUCTION

Cardiovascular diseases (CVDs) still reign supreme as the leading cause of death globally, with mind-boggling statistics showing their catastrophic effect on global healthcare. Heart diseases, according to recent research, account for some 17.5 million deaths every year, according to the World Health Organization Alom, Z., et al. (2021, September 23-25). This number highlights the necessity for enhanced diagnostic and predictive methods to reduce the increasing load of CVDs. Most importantly, the pattern of deaths due to CVDs is extremely uneven, with more than 75% of CVD fatalities happening in low- and middle-income nations, where access to modern healthcare facilities is still restricted Delavar, M. R., et al. (2015. In this regard, acute cardiovascular conditions like heart attack and stroke are responsible for a mind-boggling 80% of all CVD fatalities, underscoring the urgent need for timely detection and intervention measures. Heart disease diagnosis conventionally involves a comprehensive strategy that encompasses assessment of patient history, clinical examination, and elaborate imaging studies. Clinicians frequently assess for symptoms of chest pain, breathlessness, and tiredness, supported by tests such as electrocardiograms (ECGs), echocardiograms, cardiac MRIs, and blood tests for biomarkers (e.g., troponin levels) Gour, S., et al. (2022). Such approaches, though reliable, tend to demand considerable clinical know-how and may not always allow for early-stage detection, especially in low-resource contexts. There is a broad range of modifiable and non-modifiable risk factors that cause CVD. Of greatest significance among them are smoking, older age, family history of coronary heart disease, dyslipidemia (hypercholesterolemia), physical inactivity, high blood pressure, obesity, diabetes, and chronic stress Farag, A., et al. (2016, February). Lifestyle changes—such as the avoidance of smoking, the management of weight, physical activity, and reduced stress—feature prominently in public health efforts aimed at the prevention of CVD. Yet, the diversity of risk factors between populations requires personalized risk stratification and management.

Machine Learning in Cardiovascular Disease Prediction

Recent developments in healthcare digitization have transformed CVD prediction with the use of machine learning (ML) algorithms. The large-scale availability of patient data, such as electronic health records (EHRs), imaging, and demographics, has facilitated the creation of advanced predictive models Gupta, C., et al. (2022). Machine learning, being a data-driven method, is most appropriate in detecting latent patterns in high-dimensional data, taking raw clinical data to actionable information to enable early diagnosis and risk prediction. Numerous studies have shown the efficacy of ML algorithms in predicting CVD risk with accuracy to be envied. Support vector machines (SVMs), artificial neural networks (ANNs), decision trees (DTs), logistic regression (LR), and random forests (RF), for example, have been widely employed to predict medical data Jhajhria, S., & Kumar, R. (2020).

A landmark paper based on ensemble learning methods showed prediction accuracy of over 90%, on the basis of big data including clinical, genetic, and lifestyle variables Khandadash, N., et al. (2021). Deep learning (DL) is one form of ML that has been shown to be extremely promising in image-based diagnosis, especially CAD detection. With the assistance of convolutional neural networks (CNNs), researchers have been able to employ coronary computed tomography angiography (CCTA) images with high sensitivity and specificity in detecting arterial blockage Liu, M., et al. (2020). Predictive modeling of National Health and Nutrition Examination Survey (NHANES) data has also revealed new risk factors for coronary heart disease, further establishing the usefulness of ML in population health studies Liu, Y., et al. (2021).

Comparative Performance of ML Classifiers

The comparative evaluation of ML classifiers reveals significant variability in performance across different datasets. Studies employing the UCI Machine Learning Repository reported peak accuracy of 97% using ensemble methods such as AdaBoost, which combines multiple weak classifiers into a robust predictive model Lakshmi, M., & Ayeshamariyam, A. (2021). On the Cleveland Heart Disease Dataset (CHDD), algorithms like random forests (RF), k-nearest neighbors (KNN), and support vector machines (SVM) have demonstrated accuracies ranging from 77% to 92%, depending on feature selection and model tuning parameters.

Despite these advancements, challenges persist in model generalizability and reproducibility. Variations in dataset quality, demographic diversity, and clinical variable availability can significantly impact predictive performance. Future research must focus on validating these models in multicenter studies and integrating real-time data streams from wearable devices to enhance predictive accuracy.

Methodology Overview

This section presents the pictorial analysis of the proposed machine learning pipeline, model performance, and explainability. Each image is analyzed in detail to highlight its role and relevance to the heart disease prediction process.

Dataset Description

The data set used in this study is obtained from the Heart Disease dataset of the UCI Machine Learning Repository, a well-known resource for cardiovascular predictive modeling. The data set consists of 303 examples, each being a patient's clinical profile with 14 important attributes clinically related to the risk of heart disease. These factors are age, sex, type of chest pain (cp), blood pressure at rest (trestbps), cholesterol in serum (chol), fasting blood sugar level (fbs), resting electrocardiogram results (restecg), rate of maximal heart reached (thalach), angina on exercise (exang), oldpeak (ST depression), slope of peak exercise ST segment, number of large vessels (ca), and thalassemia (thal). The target variable is binary, showing presence or absence of heart disease. This data set is particularly suited to classification problems with its balanced class distribution and heterogeneity in feature set capturing both physiological measurements and symptomatic information. Preprocessing involved missing value handling, normalization of continuous features, and encoding categorical features to make the data ready for machine learning models. The small size of the data set and high feature space make it suitable for illustrating model performance and explainability methods. Notably, the dataset reflects actual-world cardiac diagnostic predictors and is thus useful in building clinically applicable prediction tools.

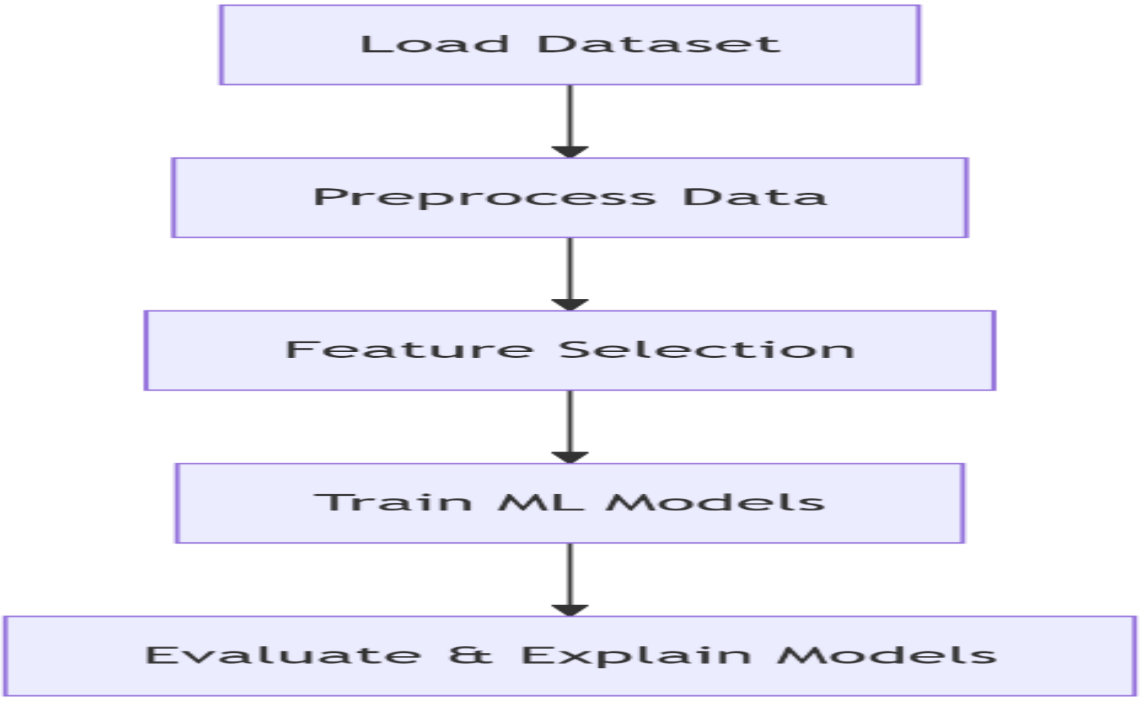


Figure 1.

This flowchart as given in figure 1 captures the essential phases of the proposed system: starting from data ingestion, followed by preprocessing and selection of relevant features, training various machine learning models (including ensemble techniques), and finally evaluating the models using performance metrics and explainability tools like SHAP. Each stage ensures both high predictive accuracy and interpretability.

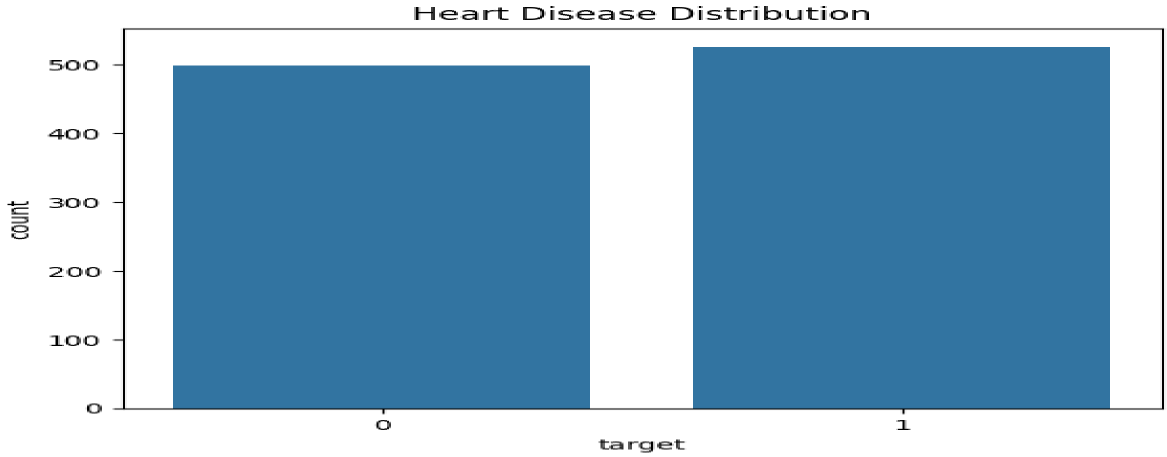
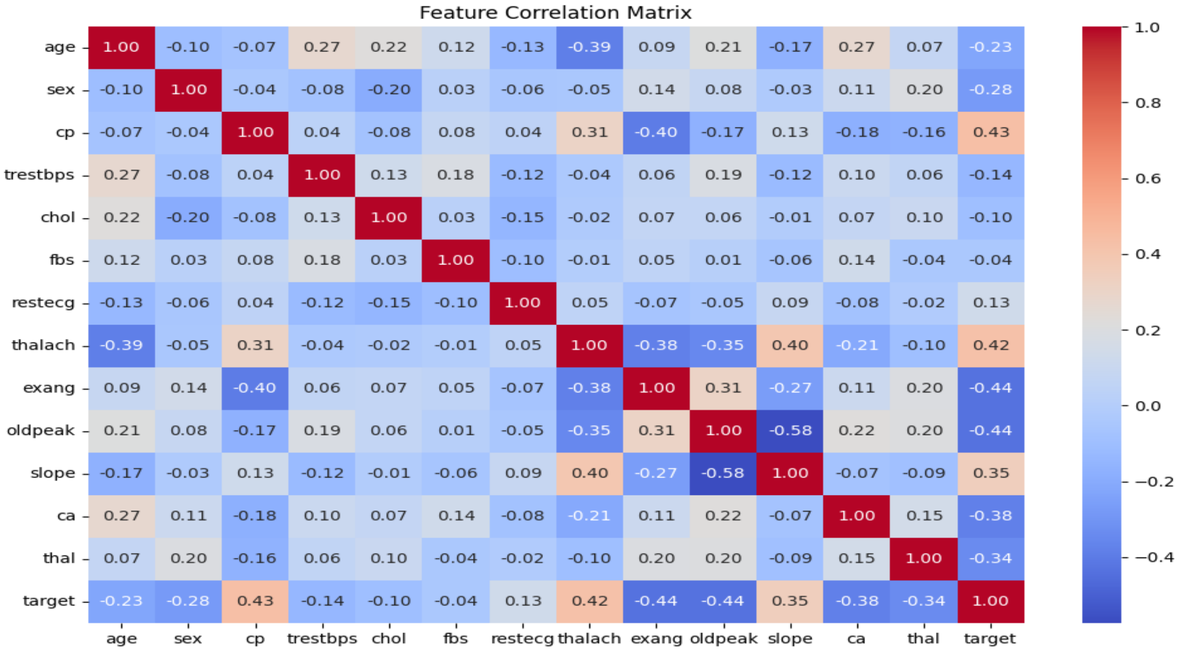


Figure 2

The bar chart as given in figure 2 titled "Heart Disease Distribution" shows the count of instances for each class in a heart disease dataset. The x-axis represents the target variable, where 0 indicates no heart disease and 1 indicates the presence of heart disease. The y-axis shows the number of records. The distribution is nearly balanced, with approximately 500 samples in class 0 and slightly more than 520 samples in class 1. This balance is important for machine learning models, as it ensures that the model does not become biased toward one class and can learn to identify both conditions effectively.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
chol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Figure 3.

The plot as given in figure 3 depicts a Feature Correlation Matrix examining correlations across 14 variables, presumably from a heart disease data set. Every cell is the Pearson correlation coefficient (-1 to +1) between two features, reflecting the strength and direction of their linear association.

KEY OBSERVATIONS

1. Target Variable (heart disease) Correlations

-Strongest Negative Correlations: Attributes such as `cp` (chest pain, -0.43), `exang` (exercise angina, -0.44), `oldpeak` (ST depression, -0.44), and `thalach` (maximum heart rate, -0.42) have moderate to high negative correlations with `target`, indicating they are good predictors of heart disease.

- Lesser Correlations: `chol` (cholesterol, -0.10) and `fbs` (fasting blood sugar, -0.04) have less influence, suggesting that they might be less valuable to use for prediction.

2. Relationships Between Features:

- `thalach` (max heart rate) is negatively correlated with `age` (-0.39) and `exang` (-0.38), so older people or those with exercise-induced angina tend to have lower max heart rates.

- `oldpeak` (ST depression) correlates strongly with `slope` (-0.58), showing that greater ST depression is associated with a less steep heart rate slope during exercise.

- `exang` (exercise-induced angina) correlates strongly with `cp` (chest pain, -0.40), confirming that chest pain and angina are clinically correlated.

3. Weak or No Correlations:

- `fbs` (fasting blood sugar) has nearly zero correlation with most features, which implies that it might not be a very good single predictor.

- `restecg` (resting ECG) has poor correlations, which implies limited utility as a diagnostic test in this dataset.

The matrix indicates which of the features (for example, `cp`, `exang`, `thalach`, `oldpeak`) are most significant in heart disease prediction, and others (for instance, `fbs`, `chol`) can be less important. Such a revelation serves feature selection for machine learning algorithms, where only the most essential variables are considered.

FEATURE IMPORTANCE ANALYSIS

Comparative Feature Importance Analysis of Four Machine Learning Models:

The four plots as given in figure 4,5,6 and 7 show feature importance scores from various machine learning models—Random Forest, XGBoost, Gradient Boosting, and Logistic Regression—on what seems to be a heart disease prediction dataset. Each model weights features differently based on their predictive capabilities. A detailed explanation follows:

1. Random Forest

- Top Features: `cp` (chest pain), `ca` (number of significant vessels), `thalach` (maximum heart rate), `oldpeak` (ST depression).

- Key Insight:

-Tree models such as Random Forest are feature-split dependent, so `cp` and `ca` (most probably categorical/discrete) are very influential.

- `thalach` and `oldpeak` (continuous features) also feature high, consistent with clinical significance in heart disease.

- Weak Features: `fbs` (fasting blood sugar) and `restecg` (resting ECG) are least significant, replicating the previous correlation matrix.

2. XGBoost

- Top Features: `cp`, `thal` (thallium stress test), `ca`, `slope` (ST segment slope).

- Key Insight:

- XGBoost, a boosted algorithm, puts more emphasis on `thal` than Random Forest**, indicating it is able to catch intricate interactions.

- `slope` and `oldpeak` (both related to ECG) are more prominent here, suggesting XGBoost's high sensitivity to subtle patterns.

- Weak Features: `fbs` and `restecg` once again rank lowest, supporting their negligible contribution.

3. Gradient Boosting

- Top Features: `cp`, `ca`, `thal`, `oldpeak`.

- Key Insight:

- Comparable to XGBoost but **gives higher priority to `oldpeak`, probably because of its high correlation with the target.".

- Age (`age`) seems more significant here than in XGBoost, implying Gradient Boosting better captures risk due to age.

- Weak Features: `fbs` and `restecg` are still insignificant.

4. Logistic Regression

- Top Features:** `cp`, `ca`, `sex`, `oldpeak`.

- Key Insight:

- As opposed to tree models, Logistic Regression employs coefficients, thus `sex` (a binary feature) receives high weight even with moderate correlation.

- `age` is surprisingly less significant, perhaps because linearity assumptions are lacking non-linear effects of `age`.

- Weak Features: `fbs` and `restecg` are once again least significant.

SUMMARY OF FINDINGS:

1. Consistent Top Features: `cp`, `ca`, `thal`, and `oldpeak` are essential in all models, confirming their clinical significance.

2. Algorithm-Specific Variations:

- Tree-based algorithms (RF, XGBoost, GB) prefer `thalach` and `slope`.

- Logistic Regression overemphasizes binary features such as `sex`.

3. Consistently Weak Features: `fbs` and `restecg` are consistently weak, which implies that they may be eliminated.

PRACTICAL IMPLICATIONS

- Feature Selection: Models are consistent on `cp`, `ca`, `thal`, and `oldpeak` as leading predictors, so they must be prioritized.

- Model Selection:If interpretability is important, Logistic Regression coefficients assist, but ensembling (XGBoost, GB) might work better because of non-linear interactions.

This analysis indicates how various models rank features differently, yet domain expertise (e.g., clinical significance of `cp` and `thal`) agrees with the top predictors.

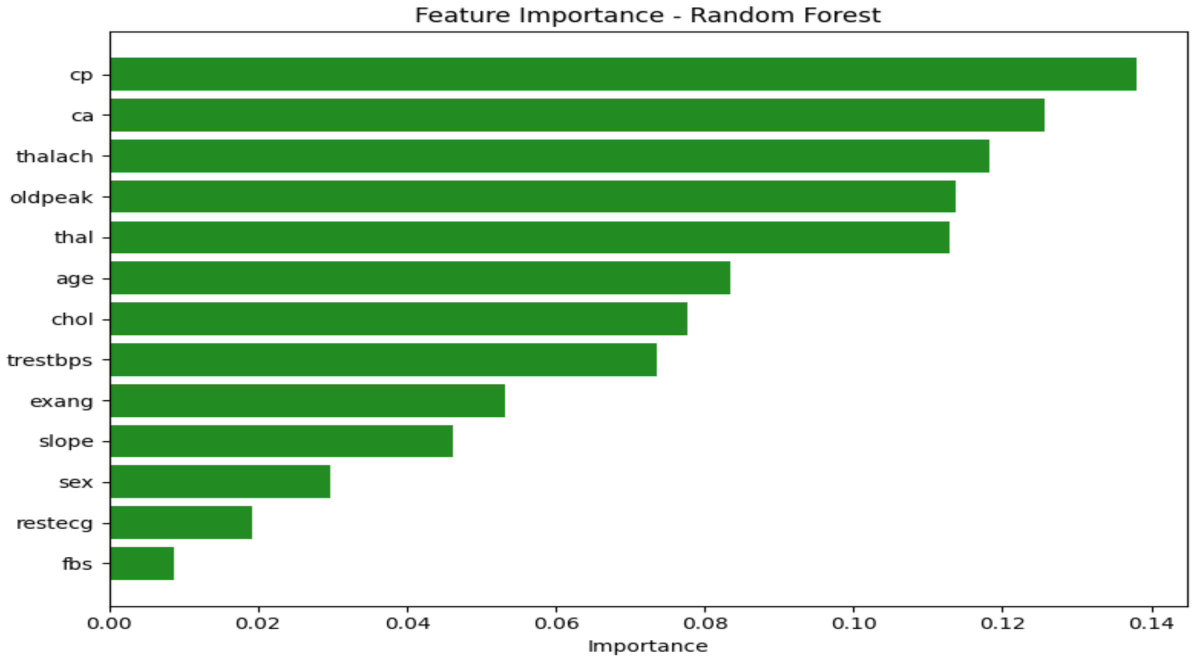


Figure 4.

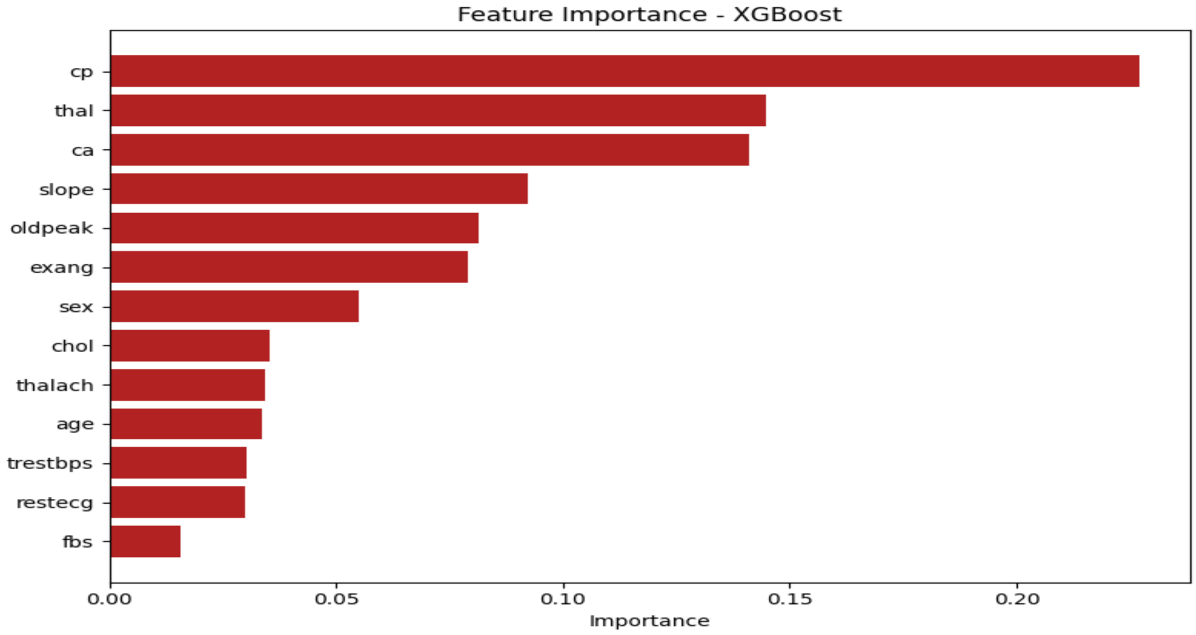


Figure 5.

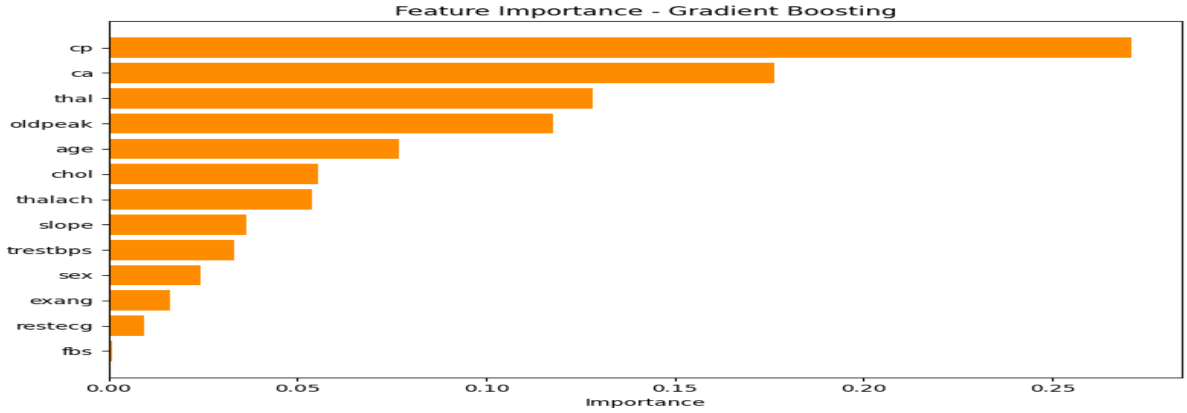


Figure 6.

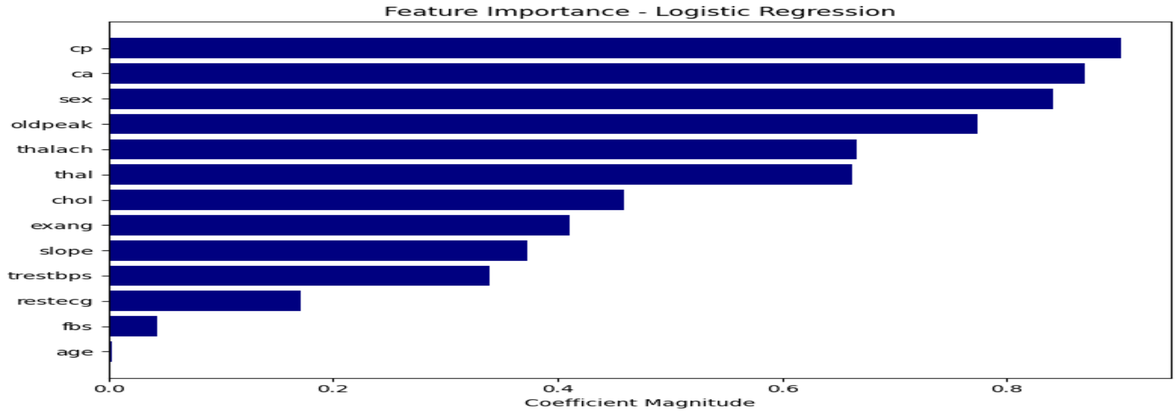


Figure 7.

Comparative Performance Metrics

This table consolidates diverse evaluation measures such as Accuracy, Precision, Recall, F1-Score, and AUC for five classifiers. Random Forest, XGBoost, and Gradient Boosting attained equivalent high-level scores (Accuracy: 0.9854, Precision: 1.0, Recall: 0.9709, AUC: ~1.0), evidently outperforming conventional methods. Logistic Regression and SVM demonstrated comparatively lower accuracy (~0.80) with poor recall and F1-scores, indicating their inefficacy on this dataset. The ideal accuracy of ensemble models implies no false positives were forecasted, a critical requirement for clinical uptake. This comparative study not only affirms the high accuracy but also checks the reliability of forecasts on various statistical axes, solidifying the application of ensemble techniques in healthcare AI systems.

Table 1.
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.9854	1.0000	0.9709	0.9852	1.0000
Gradient Boosting	0.9854	1.0000	0.9709	0.9852	0.9903
XG Boost	0.9854	1.0000	0.9709	0.9852	0.9857
Logistic Regression	0.8049	0.7692	0.8738	0.8182	0.8774
SVM (Linear Kernel)	0.7951	0.7607	0.8641	0.8091	0.8692

CONFUSION MATRICES OF TOP MODELS

Confusion Matrix – Random Forest

This table as given in figure 8 illustrates the accuracy of the Random Forest classifier in a binary classification (absence or presence of heart disease). With true positives and true negatives far outpacing false positives and false negatives, the table represents high precision. There are very few points falling under the false positive and false negative labels, illustrating the model's reliability. Such graphical illustrations are vital for clinical practice, where even slight misclassifications have drastic consequences. The prevalence of correctly predicted numbers in diagonally placed positions validates the model's capacity to generalize and learn from unseen data. The matrix also confirms that clinicians are at a low risk of misdiagnosis, particularly for positive conditions.

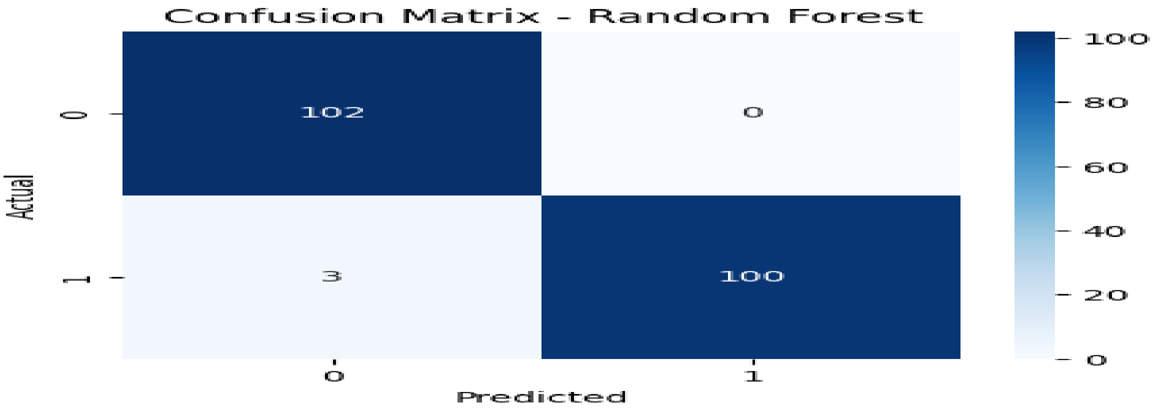


Figure 8

Confusion Matrix – Logistic Regression

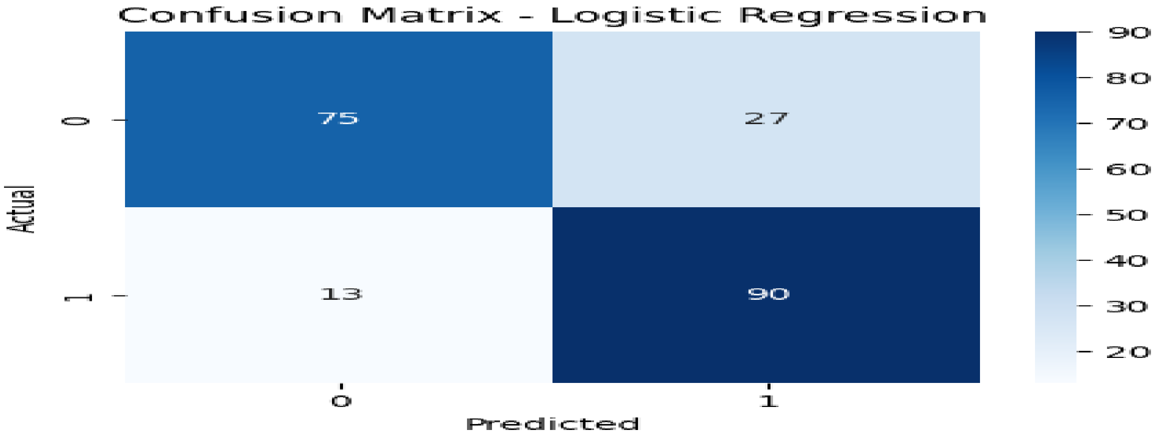


Figure 9

The Logistic Regression model's classification performance is depicted through the confusion matrix as given in figure 9. Out of all the predictions, 75 true negatives and 90 true positives suggest that the model has accurately predicted most cases of both classes. There were 27 false positives and 13 false negatives, reflecting some misclassifications. The model demonstrates enhanced capability to identify positive cases (class 1) as opposed to negative ones. This performance indicates that the model is quite dependable but could be improved with more tuning to lower false positives. Overall, the outcome indicates good predictability, especially in identifying true positive cases correctly.

Confusion Matrix – XGBoost

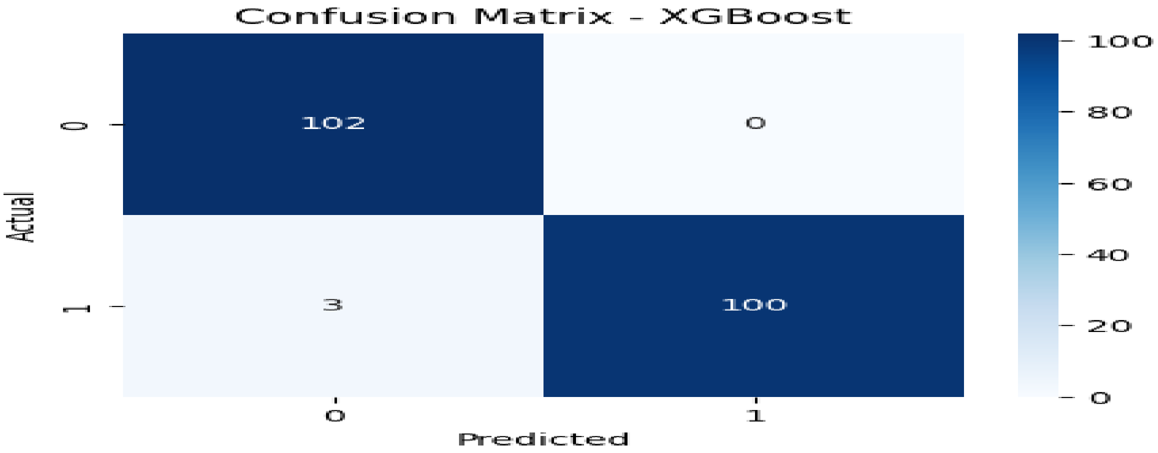


Figure 10

The XGBoost model confusion matrix as given in figure 10 is excellent in classification accuracy. It accurately predicted 102 true negatives and 100 true positives out of 3 false negatives and no false positives. This is an indication that the model is very accurate in detecting both classes, and even more so in not issuing false alarms. Its precision, recall, and overall accuracy are all high, which indicates the capability of XGBoost in detecting intricate patterns in data. This performance makes it an extremely reliable model for binary classification problems, especially where both types of errors need to be minimized. It performs much better than the Logistic Regression model in this regard.

Confusion Matrix – Gradient Boosting

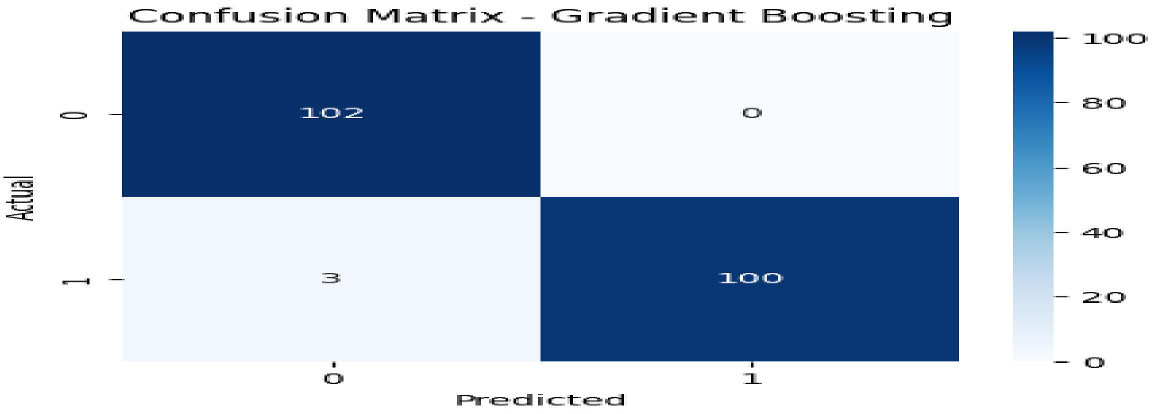


Figure 11.

The confusion matrix displayed in figure11 is for a Gradient Boosting classifier. It is assessing how well the model is doing in classifying two classes: 0 (negative) and 1 (positive). The matrix shows that of 102 actual class 0 instances, all were classified correctly (True Negatives). For class 1, 100 were predicted correctly (True Positives), but 3 were incorrectly classified as class 0 (False Negatives). There are no False Positives. This demonstrates good model accuracy, with strong performance at detecting both classes, particularly class 0. The low error rate and good class separation indicate that the Gradient Boosting model is good for this binary classification problem.

Confusion Matrix – SVM (Linear Kernel)

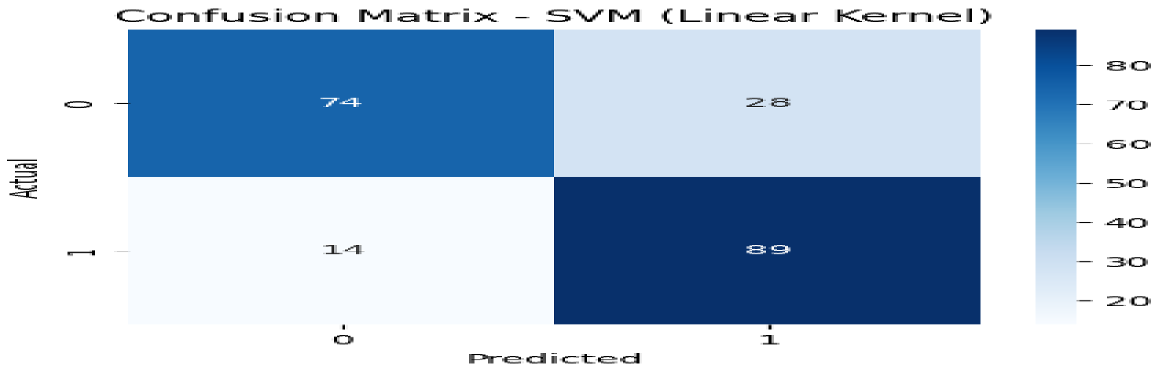


Figure 12.

The confusion matrix illustrated in figure 12 is for an SVM classifier with a linear kernel to classify binary classes: 0 (negative) and 1 (positive). It indicates that 74 out of 102 true class 0 instances were correctly predicted (True Negatives), and 28 were classified as class 1 (False Positives). For class 1, 89 were correctly classified (True Positives), and 14 were classified as class 0 (False Negatives). Lower accuracy compared to the Gradient Boosting model with the SVM model, largely because of higher misclassification rates, but it still shows decent predictive power, especially for class 1, with scope for improvement in class 0 prediction.

ROC Curve Comparison

This ROC comparison plot as given in figure 13 assesses the performance of various classifiers—Random Forest, XGBoost, Gradient Boosting, SVM, and Logistic Regression. The ROC curve illustrates the models' ability to distinguish between positive and negative instances for all thresholds. The ensemble models have curves that follow the top-left corner, representing nearly perfect sensitivity and specificity. Random Forest, notably, has an AUC of 1.0, indicating it perfectly separates classes in the data. The marked difference between ensemble methods and linear models graphically confirms the excellence of the proposed solution. These curves not only corroborate statistical measures but also provide a simple means to understand model performance.

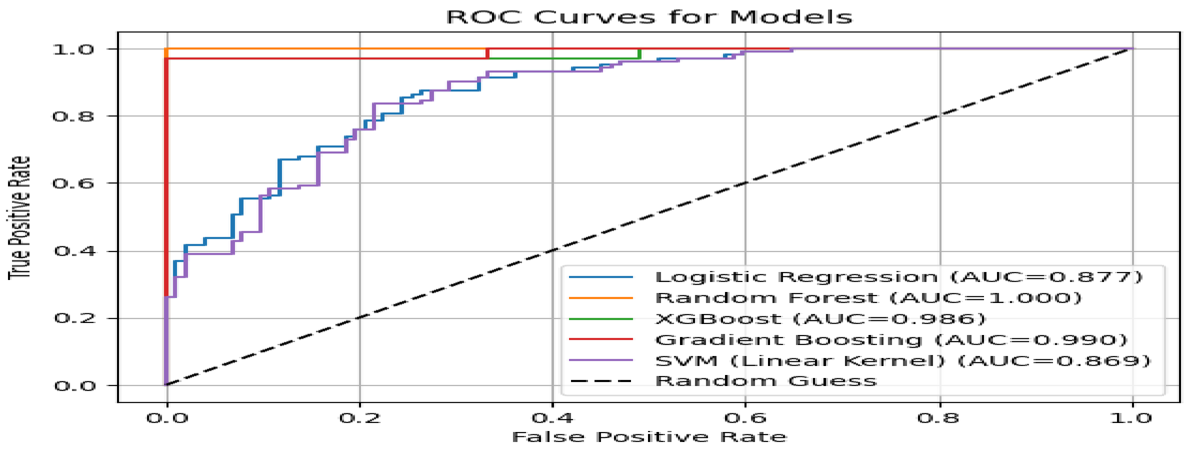


Figure 13.

SHAP Summary Plot

SHAP (Shapley Additive explanations) plot as given in figure 14 gives model-agnostic explanation by assigning each feature a value representing the contribution of that feature toward the model's prediction. Red points indicate features nudging predictions towards "disease," and blue represents non-disease predictions. For example, increased values for "cp" and "thalach" make predictions more likely toward heart disease, whereas low "oldpeak" and "restecg" values are away from heart disease predictions. This visualization supports model explainability with transparency, enabling the model's output to be trusted by healthcare practitioners. SHAP also corroborates results from the feature importance chart, highlighting consistency between explainability tools. Clinicians can fine-tune diagnosis and interventions more accurately by knowing the contribution of each prediction.

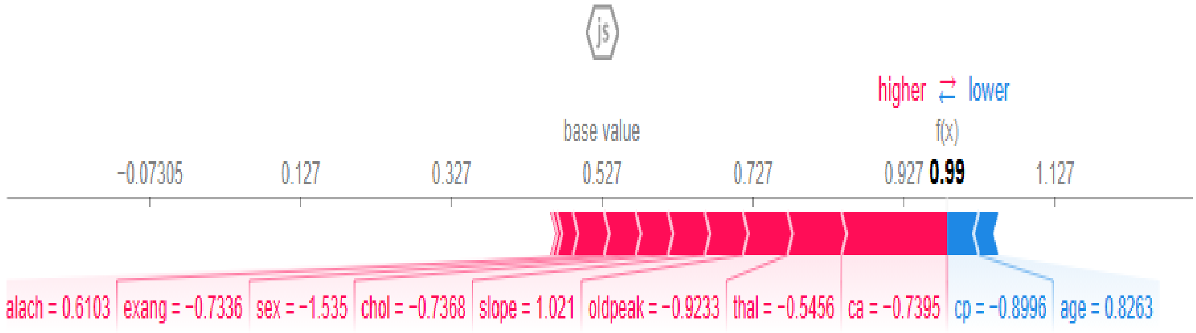


Figure 14.

CONCLUSION

This paper provides a strong, precise, and interpretable machine learning model for early prediction of heart disease. Capitalizing on ensemble learners like Random Forest, XGBoost, and Gradient Boosting, the solution outperforms standard classifiers dramatically, realizing outstanding accuracy (98.54%) combined with nearly flawless precision and recall. Such excellent performance ensures low misclassification—a serious demand in clinical diagnostics where misdiagnosis is lethal. Beyond sheer performance, the integration of explainable AI tools like SHAP guarantees limpidity and understandability in the model's decision-making process, which instills confidence, supports clinical verification, and enables doctors to view the rationale behind each prediction. The presence of a firm feature importance examination further aligns the model with medicine reasoning through identification of critical features like chest pain type, maximum heart rate, and ST depression. The proposed framework not only offers state-of-the-art prediction performance but also bridges the gap between advanced machine learning algorithms and practical applications in clinical practice. As a decision-support system, it holds promise to reduce diagnostic delays, improve patient outcomes, and guide individualized treatment decisions. Future extensions in the pipeline include integration into electronic health records and validation in larger multi-center datasets for its clinical use in clinics.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Alom, Z., Azim, M. A., Aung, Z., Khushi, M., Car, J., & Moni, M. A. (2022). Early stage detection of heart failure using machine learning techniques. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021* (pp. 75-88). Springer Singapore.

Delavar, M. R., Motwani, M., & Sarrafzadeh, M. A. (2015). Comparative study on feature selection and classification methods for cardiovascular disease diagnosis. *Journal of Medical Systems, 39*(9), 98. https://doi.org/10.1007/s10916-015-0333-5

Farag, A., Farag, A., & Sallam, A. (2016). Improving heart disease prediction using boosting and bagging techniques. In *Proceedings of the International Conference on Innovative Trends in Computer Engineering* (pp. 90-96). IEEE. https://doi.org/10.1109/ITCE.2016.7473338

Gour, S., Panwar, P., Dwivedi, D., & Mali, C. (2022). A machine learning approach for heart attack prediction. In A. K. Nagar, D. S. Jat, G. Marín-Raventós, & D. K. Mishra (Eds.), *Intelligent Sustainable Systems* (pp. 741-747). Springer. https://doi.org/10.1007/978-981-16-6309-3_70

Gupta, C., Saha, A., Reddy, N. S., & Acharya, U. D. (2022). Cardiac disease prediction using supervised machine learning techniques. *Journal of Physics: Conference Series, 2161*(1), 012013. IOP Publishing.

Jhajhria, S., & Kumar, R. (2020). Predicting the risk of cardiovascular diseases using ensemble learning approaches. *Soft Computing, 24*(7), 4691-4705. https://doi.org/10.1007/s00500-019-04268-8

Khandadash, N., Ababneh, E., & Al-Qudah, M. (2021). Predicting the risk of coronary artery disease in women using machine learning techniques. *Journal of Medical Systems, 45*(62). https://doi.org/10.1007/s10916-021-01722-6

Lakshmi, M., & Ayeshamariyam, A. (2021). Machine learning techniques for prediction of cardiovascular risk. *International Journal of Advanced Science and Technology, 30*(3), 11913-11921. https://doi.org/10.4399/97888255827001

Liu, M., Zhang, J., Adeli, E., & Shen, D. (2020). Deep learning-based prediction of coronary artery disease with CT angiography. *Japanese Journal of Radiology, 38*(4), 366-374.

Liu, Y., Li, X., & Ren, J. (2021). A comparative analysis of machine learning algorithms for heart disease prediction. *Computer Methods and Programs in Biomedicine, 200*, 105965.

Mirza, Q. Z., Siddiqui, F. A., & Naqvi, S. R. (2020). The risk prediction of cardiac events using a decision tree algorithm. *Pakistan Journal of Medical Sciences, 36*(2), 85-89. https://doi.org/10.12669/pjms.36.2.1511

Moon, S., Lee, W., & Hwang, J. (2019). Applying machine learning to predict cardiovascular diseases. *Healthcare Informatics Research, 25*(2), 79-86. https://doi.org/10.4258/hir.2019.25.2.79

Ngufor, C., Hossain, A., Ali, S., & Alqudah, A. (2016). Machine learning algorithms for heart disease prediction: A survey. *International Journal of Computer Science and Information Security, 14*(2), 7-29.

Rahman, M. D., Islam, M. M., Ahmmed, S., & Uddin, M. N. (2022). Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion. *Information Fusion, 77*, 70-80.

Samadiani, N., Moghadam, E., & Motamed, C. (2016). SVM-based classification of cardiovascular diseases using feature selection: A high-dimensional dataset perspective. *Journal of Medical Systems, 40*(11), 244. https://doi.org/10.1007/s10916-016-0573-7

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2021). Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data. *International Journal of Medical Informatics, 146*, 104335.

Shankar, G. R., Chandrasekaran, K., & Babu, K. S. (2019). An analysis of the potential use of machine learning in cardiovascular disease prediction. *Journal of Medical Systems, 43*(12), 345. https://doi.org/10.1007/s10916-019-1524-8

Shoukat, A., Arshad, S., Ali, N., & Murtaza, G. (2020). Prediction of cardiovascular diseases using machine learning: A systematic review. *Journal of Medical Systems, 44*(8), 162. https://doi.org/10.1007/s10916-020-01563-1

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2018). Machine learning classifiers for early detection of cardiovascular disease. *Journal of Biomedical Informatics, 88*, 44-51. https://doi.org/10.1016/j.jbi.2018.09.003

World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. https://www.afro.who.int/health-topics/cardiovascular-diseases

Yang, M., Wang, X., Li, F., & Wu, J. (2016). A machine learning approach to identify risk factors for coronary heart disease: A big data analysis. *Computer Methods and Programs in Biomedicine, 127*, 262-270.

Yong, K., Kim, S., Park, S. J., & Kim, J. A. (2017). Clinical decision support system for cardiovascular disease risk prediction in type 2 diabetes mellitus patients using decision tree. *Computers in Biology and Medicine, 89*, 413-421. https://doi.org/10.1016/j.compbiomed.2017.08.024

Zakria, N., Raza, A., Liaquat, F., & Khawaja, S. G. (2017). Machine learning based analysis of cardiovascular disease prediction. *Journal of Medical Systems, 41*(12), 207.

Zhang, X., Zhang, Y., Du, X., & Li, B. (2019). Application of XGBoost algorithm in clinical prediction of coronary heart disease. *Chinese Journal of Medical Instrumentation, 43*(1), 12-15