Comprehensive Journal of Science

Volume (10), Issue (37), (NOV. 2025)

SICST2025, www.sicst.ly

ISSN: 3014-6266

مجلة العلوم الشاملة

عدد خاص بالمؤتمر الدولي الثالث للعلوم والتقنية

المجلد (10)، العدد (37)، (نوفمبر2025)

ردمد: 3014-6266

Enhancing Heart Disease Prediction through Effective Handling of Data Imbalances and the Use of Ensemble Learning Techniques.

Noura Marie Elshiebani (1)

Higher Institute of Science and Technology - Suluq

noraal-shaibani@hicps.edu.ly

Geber Khalifa Geber (2)

Higher Institute of Science and Technology - Suluq

geber@hicps.edu.ly

Mostafa Nser brka (3)

Higher Institute of Science and Technology, Al Shomokh

mostafazanate86@shomokh.edu.ly

Abraheem Mohammed sulayman Alsubayh(4)

Faculty of Arts and Sciences, University of Benghazi, Solouq, Libya

abraheem.alsubayhay@uob.edu.ly

Received: 30-09-2025; Revised: 10-10-2025; Accepted: 31-10-2025; Published: 25-11-2025

Astract

This study investigates the prediction of heart disease severity by utilising the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The research commenced with extensive data preprocessing, such as handling missing values and applying one-hot encoding to categorical variables. Several machine learning models were tested, and the combination of LightGBM with SMOTE produced the highest accuracy (0.6739) and ROC AUC (0.8941).

To further improve predictive performance, a stacking ensemble approach was applied, combining the strengths of several machine learning models. Age and exercise-induced angina emerged as critical predictors, offering insights for early detection and better management of heart disease. These results highlight the importance of these variables in reducing heart disease severity and point toward possible improvements in patient care. The analysis utilised Python in Google Colab, leveraging its extensive libraries and tools to achieve precise results.

Keywords: Heart Disease Prediction; Cardiovascular Disease (CVD); Machine Learning; Feature Selection; Class Imbalance; SMOTE; Ensemble Learning; Stacking.

Introduction

Heart disease is one of the leading causes of death worldwide and accounted for approximately 32% of total deaths in 2019. The number clearly exhibits the necessity of effective mechanisms in prediction; then, CVD is a growing concern regarding the health of young people. For example, 1,122 university students exhibited good knowledge regarding modifiable risk factors, such as smoking (0.853) and hypertension (0.858), but less than 40% of them were aware of non-modifiable risks, such as age and genetics.

The prime modifiable risk factors are hypertension, smoking, diabetes, obesity and inactivity that can be controlled with changes in lifestyle. The non-modifiable factors such as age, sex and family history also increase the risk of CVD. Drinking and stress increase the risk, indicating the necessity of better knowledge regarding cardiovascular health. In predictive modelling, those risk factors are translated into features that influence classification output. Even so, all the features may not be equally contributing to the accuracy of prediction, which is the reason that feature selection techniques become vital. Prioritising the most informative predictors and possessing knowledge of them can enhance model performance as well as produce more transparent information regarding the determinants of cardiovascular risk. [1]

The study applies mixed statistical techniques that include Recursive Feature Elimination (RFE). Pearson correlation, chi-square test and analysis of variance (ANOVA) in studying biomarkers for cardiovascular disease CVD and establishing significance through p-value testing. Despite the effectiveness of the method in establishing the significant correlation of biomarkers and CVD, loopholes lie in a comprehensive understanding of the biological process involved and the probable role played by unadjusted variables. [2] The researcher conducting the study is H. Abdelhalim, who recognises the necessity for additional research to bridge such limitations and increase the precision of CVD diagnostics

Bhatt CM et al. bring out the importance of preprocessing and feature effect on accuracy in models such as Bhatt et al. 2023, who managed to shrink their dataset to 59,000 rows and 11 features, making their machine learning models more efficient in predicting heart disease. [3] Sachan Almaghrabi Yang and Xu's 2022 [4] research brings out the importance of data preprocessing techniques such as transformation, cleansing, and balancing for enhancing the quality of heart disease prediction datasets. It used the ANOVA-F test in feature selection, and significant medical features such as age, hypertension, glucose and blood pressure were found to be crucial in enhancing the accuracy of prediction. Apart from that, the study brings out eliminating unnecessary features such as gender and non-medical features to achieve the best model performance and maintain computational overhead low overall, pointing out the importance of selecting only relevant features to enhance the efficiency of prediction models

Choice of key variables, data collection, pre-processing of data, handling missing values, Cleaning of data and data normalisation are all critical steps in the research methodology proposed to predict heart disease. Jindal et al. (2021) expound on how attribute selection and preprocessing play a crucial role in determining the accuracy of the models utilised in heart disease prediction, where KNN and logistic regression were accurate to 0.885. [5] Dr Manivannan's research examines the use of synthetic data by CTGAN to improve machine learning classifiers in intrusion detection with imbalanced datasets. Findings indicate classifiers were more accurate and achieved higher metrics with the use of CTGANSamp compared to utilising original datasets ORG and random sampling. RandomSamp Though there are gaps in the long-term effect of synthetic data on the generalisation of models and future application of CTGAN in other fields, it also exposes the Matthews correlation coefficient MCC as a performance metric for imbalanced datasets, with improved results yielded using CTGANSamp. [6]

The integration of SMOTE (Synthetic Minority Oversampling Technique) and fuzzy expert systems offers a possible way of improving the prediction of heart disease. As SMOTE generates synthetic samples for the minority class, the dataset is balanced, and the performance of machine learning algorithms improves. The utilisation of fuzzy logic in the model allows for the processing of uncertainty and imprecision in medical data, with the capability (RFE) to achieve more delicate decision-making and thus improved diagnostic yield and patient management in heart failure. [7] Apurv Garg Bhartendu Sharma, and Rijwan Khan's study utilised supervised machine learning classifiers Random Forest and K-Nearest Neighbour (K-NN) to classify an individual as having or not having heart disease based on attributes such as chest pain, cholesterol and age. The predictive accuracy of the K-NN algorithm was 0.8688, while that of the Random Forest algorithm was 0.8196. [8] Nevertheless, the study has limited use in evaluating the effectiveness of unbalanced data, and more studies utilising diverse datasets and explainable AI methods are required

Subramani et al. overcome gaps in CVD prediction by uniting machine learning and deep learning approaches, avoiding the pitfalls of traditional models prone to expecting linearity and independence of predictors. The flexibility of input selection is guaranteed by the approach via five heart disease datasets (918 samples, 11 features) and the use of GBDT with SHAP for feature selection. A two-ensemble stack with RF, LR, MLP, ET and CatBoost as base learners and LR as a meta-learner was identified to perform best with a result of an accuracy of around 96% and superior performance compared to individual models. The research also indicates the clinical utility of biomarkers such as hs-CRP, IL-6, ADP and NEFA as significant predictors. [9] There are gaps naturally, like the need for a larger and more generalizable set, and the inclusion of IoT and deep learning data to further increase the accuracy and clinical utility of predictions

The study by Moiz Qureshi, Muhammad Daniyal and Kassim Tawiah employed a random sampling design to screen 518 heart disease patients in the Lady Reading Hospital and Khyber Teaching Hospital in Khyber Pakhtunkhwa, Pakistan. Inconsistency and missing values were dropped from the data. The results indicated that the RF algorithm provided a sensitivity of 0.8611 and specificity of 0.6548, further verifying the effectiveness of their classification methods in cardiovascular disease CVD prediction [10]. Bharti et al., in their research work conducted in 2021, had also indicated a high accuracy of 0.9420 in heart disease prediction based on the use of deep learning techniques, citing the effectiveness of their method. The method entailed the utilisation of the UCI Machine Learning Heart Disease dataset with 14 attributes to be processed [11]. However, the authors noted a limitation of their study, noting that the small size of the dataset constrains the ability for deeper learning and optimisation and that more encouraging results may be obtained from larger datasets.

In another work, the authors employed a sophisticated methodology with the Cleveland heart dataset, 303 instances and 14 applicable features to classify the risk of heart disease using various machine learning classifiers (SMO and IBk), and the best accuracy of 0.8646 was obtained through the use of the SMO algorithm (i.e., A.A.A.). S.P. Sivajothi Paramasiva, M.H.N.C. and S.P. S. Pranavanand [12] The work highlighted correct selection of attributes and hyperparameter tuning for better prediction performance, but also mentioned limitations in exploring larger datasets and other algorithms for further accuracy enhancement and robustness. Umarani Nagavelli, Debabrata Samanta and Partha Chakraborty's research work employs varied machine learning methods like Naïve Bayes using a weighted approach, Support Vector Machines (SVM) and XGBoost to achieve optimal accuracy in diagnosing heart diseases. The research work uses public datasets, i.e., ECG data for experimental use, to identify the effectiveness of the models and ensure that XGBoost is best in yielding higher precision, recall and F1-measure accuracy compared to other methods. There are, however, discrepancies between acquiring more public datasets to carry out magnetocardiography (MCG) research and needing greater datasets to enhance the power and potential of the models towards generalisability in clinical usage. [13]

The work by Gárate-Escamila et al. proposed a heart disease prediction classification model with feature selection and PCA using data from 303 patients with 14 clinical features to predict heart disease. The strategy was to apply feature selection to identify the most significant features. Apply PCA for dimensionality reduction and train three classifiers: support vector machine, random forest and logistic regression with 10-fold cross-validation. The random forest classifier worked best at 0.85, followed by the other classifiers [14]. In the study of Ghulab Nabi Ahmad et al., the authors worked better compared to other studies with the utilisation of other machine learning classifiers, such as Linear Discriminant Analysis LDA and Random Forest, where Random Forest had 0.1 accuracy and Decision Tree had 0.9976. Their strategy was k-fold cross-validation, essentially measuring model performance and additionally pointing out gaps in the research work of feature selection methods, which, if further optimised, could improve predictive performance further. [15] The authors propose further research to optimise such methods for overall heart disease prediction improvement when used clinically, pointing out the advancements of their study relative to previous research

Achyut Tiwari, Aryan Chugh, and Aman Sharma created an early cardiovascular disease CVD diagnosis machine learning model using a large IEEE Data Port dataset amalgamated from several sources. The technique used was the application of a stacked ensemble classifier via the ExtraTrees Random Forest and XGBoost algorithms. The results were observed to validate 0.9234 accuracy, thus enhancing early diagnosis, which is extremely critical for the reduction of CVD-related mortality [16]. In the future, more sophisticated explainable AI XAI techniques are likely to enhance the interpretability of complicated models significantly, thus lending support to collaboration between healthcare professionals and AI systems. The research conducted by Guleria et al. (2022) outlines acute deficits in the necessity for explanation in black-box models, which have the possibility of yielding misleading outcomes in clinical decision-making. For this problem to be overcome, authors suggest a three-stage XAI process of pre-modelling, explainable modelling and post-modelling with the assurance of transparent interpretation of results. This process has been established by P.G S.A. and P.N.S [17]

Materials and Methodology

The data set is titled "Heart Disease Data" on Kaggle and was originally copied from the UCI Machine Learning Repository: Andras Jánosi, William Steinbrunn, Matthias Pfisterer, & Robert Detrano. (1989). Heart Disease [https://doi.org/10.24432/C52P4X] [18]. The research approach to heart disease prediction was to use a well-known data set that contains both numerical features (e.g., blood pressure and age) and categorical features (e.g., chest pain type and sex). As shown in the figure below, preprocessing entailed missing value management for categorical attributes by label encoding, following a two-step encoding strategy in which ordinal and binary attributes were encoded by label encoding, whereas nominal multi-class attributes were used with one-hot encoding. Target 'num' was defined as a multi-class response attribute of heart disease degree, ranging from 0 (not diseased) to 4 (most diseased). As a baseline, some of the popular machine learning classifiers like Logistic Regression, Random Forest, and XGBoost were baseline tested and then hyperparameter-tuned to achieve the best performance. For the imbalanced class, class weights were imposed on the LightGBM algorithm using the Synthetic Minority Over-sampling Technique (SMOTE), thus enhancing model accuracy. Model performances were also modelled with test accuracy, weighted F1 score, and Receiver Operating Characteristic (ROC) AUC metrics in order to estimate effectiveness. Furthermore, stacking ensemble methods were employed to wrap heterogeneous base learners like LightGBM and CatBoost and improve prediction accuracy and generalisability. This approach takes advantage of every model's natural strengths, more than a single classifier alone [19], especially when handling complex datasets and instances of class imbalance, as is the case for heart disease prediction. Thereby taking advantage of each one's strength for a strong predictive model..

Figure 1: Workflow of the proposed machine learning pipeline for heart disease prediction.

Exploratory Data Analysis

The data (Dua) has the numerical and categorical variables used in the prediction of heart disease. The numerical variables are age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), and depression of ST (oldpeak). The categorical variables are chest pain type (cp), fasting blood sugar (fbs),

resting electrocardiographic results (restecg), angina causethalassaemia (exang), slope of ST segment (slope), number of major vessels (ca), thalassaemia type (thal), data source (dataset), and sex (sex).

Preprocess to impute missing values in categorical features with an imputation strategy that internally uses label encoding to deal with non-numeric values. Following imputation, the two-step encoding process was completed: ordinal or binary categorical attributes such as sex, fbs, and exang still had label encoding, while nominal multi-class attributes such as dataset, cp, restecg, slope, thal, and ca were further one-hot encoded using pd.get_dummies(). This hybrid encoding strategy provided proper handling of

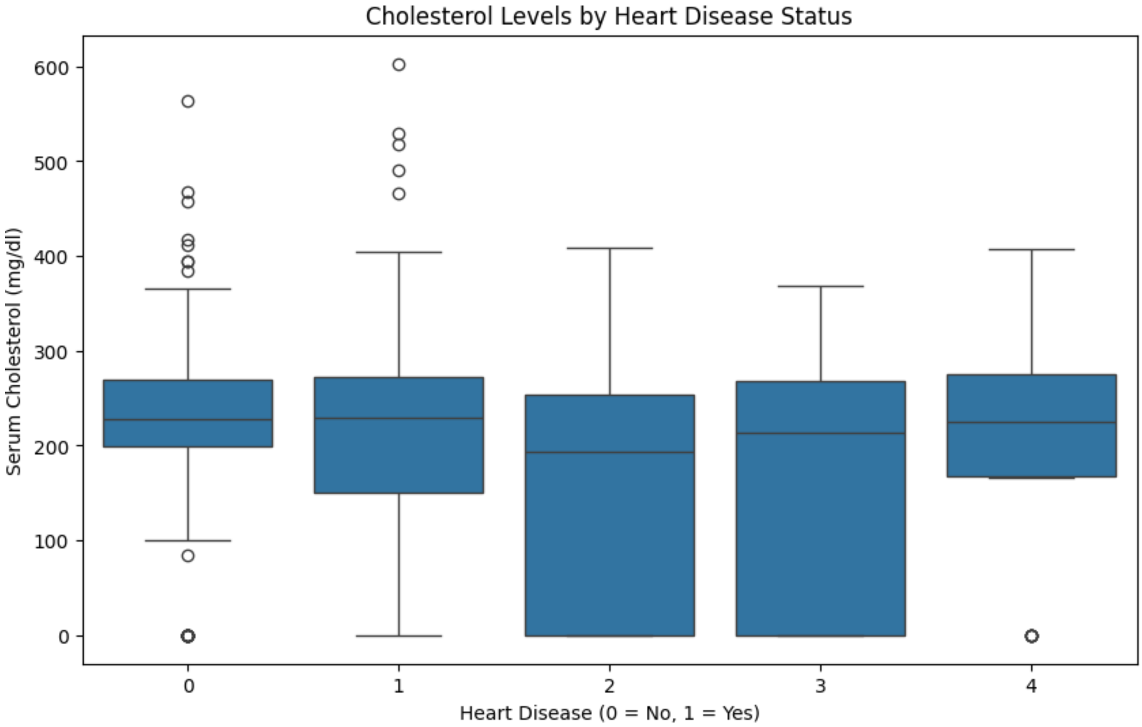
categorical features without losing feature interpretability and compatibility with machine learning models.

The target (num), severity of heart disease, was treated as a multi-class categorical variable ranging from 0 (no disease) to 4 (most severe), and this was framed as a multi-class classification problem.

Analysis of maximum heart rate (Max HR) in Figure 1 across heart disease categories (0–4) shows that individuals without heart disease (n = 0) tend to have higher

cardiovascular capacity, as shown by high levels of max HR. In contrast, subjects with increasing severity of heart disease (num = 1–4) have a progressive decline in max HR. The contrasting trend is also confirmed from a scatter plot of Age vs. Max HR in Figure 2, where ageing is associated with reduced HR and a higher occurrence of disease labels, pointing to an interacting effect of age and disease on cardiac function.

Figure 1:heart disease categories (0–4).



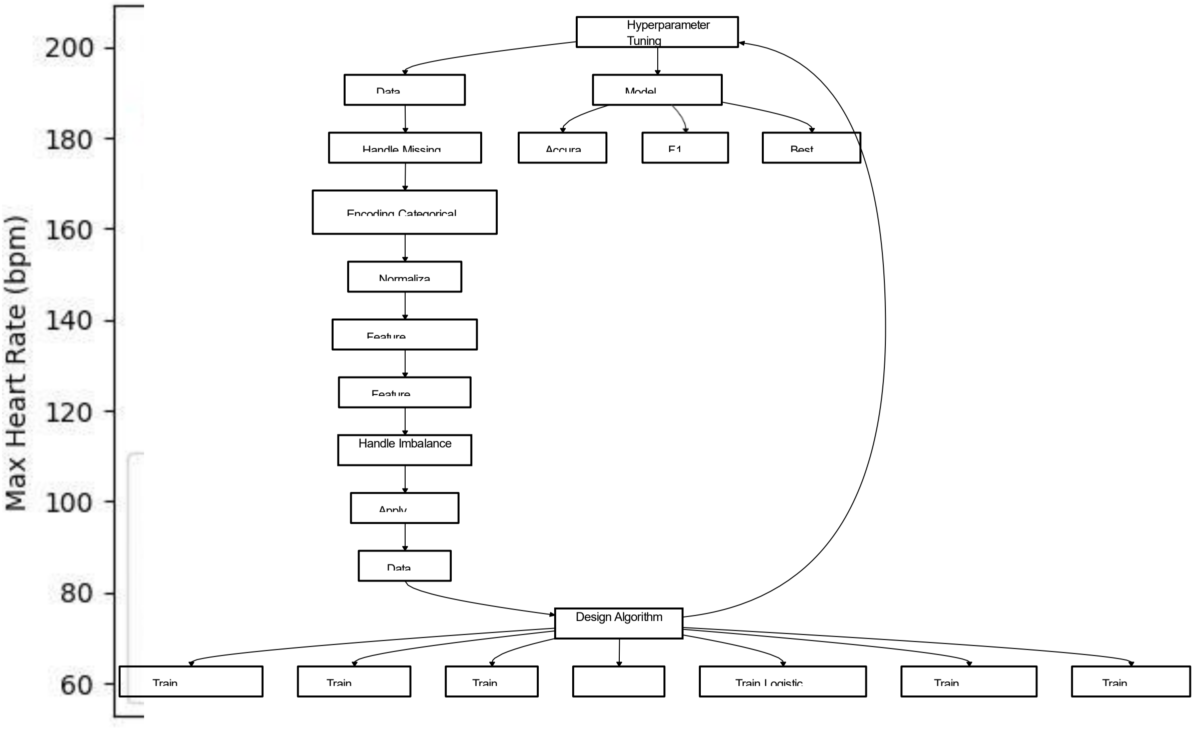


Figure 2: Age vs Max HR.

Further evidence is given by the stacked bar chart, Figure 3, showing heart disease prevalence for those with and without exercise-induced angina (exang). In those without angina, the majority are in num = 0 (no disease), but subjects with angina (exang = True) have higher numbers of serious heart conditions (particularly num = 1 and above). This trend strongly indicates that exercise-induced angina is a clinically valuable predictor both of the presence and the severity of heart disease.

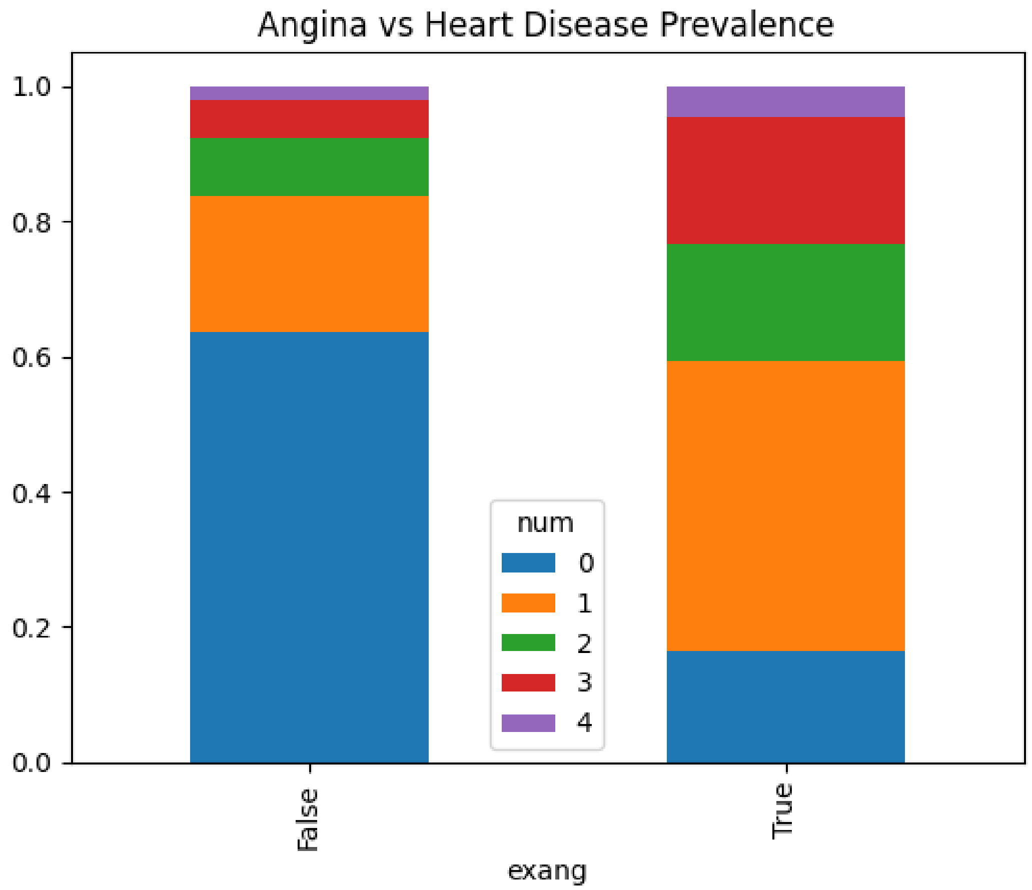


Figure 3: Angina vs Heart Disease Prevalence.

Together, these findings emphasise that Max HR, Age, and Exang are characteristics of great importance. Their joint use during model training—especially with interaction terms or non-linear transformations—can significantly improve classification accuracy. These variables need to guide feature selection, risk scoring, and decision thresholds in predictive modelling for cardiovascular risk stratification as well.

Analysis and Results

At the initial baseline modelling phase, a number of standard machine learning classifiers were compared with a stringent regime of hyperparameter optimisation and tuning.

Logistic Regression, Random Forest, XGBoost, SVC, K-Nearest Neighbours, Gradient

Boosting, AdaBoost, and HistGradientBoosting models were developed with

previously defined parameter grids. With GridSearchCV with 5-fold cross-validation, the best parameters were found in the shape of mean training accuracy. These trained models were then retrained on the full training set and tested on the unseen test set for accuracy, precision, recall, and F1 score. Results were stored in 4 and 5 DataFrames and plotted for their comparison, as can be seen in Figures 4 and 5. Tree ensemble models on trees—

i.e., XGBoost (~0.6789), GradientBoosting (~0.6762), and RandomForest (~0.6748)—were all consistently superior to their more basic counterparts in asserting their suitability in uncovering complex patterns in the data. This baseline test provided data-driven justification for further optimisation by establishing the most effective modelling standards.

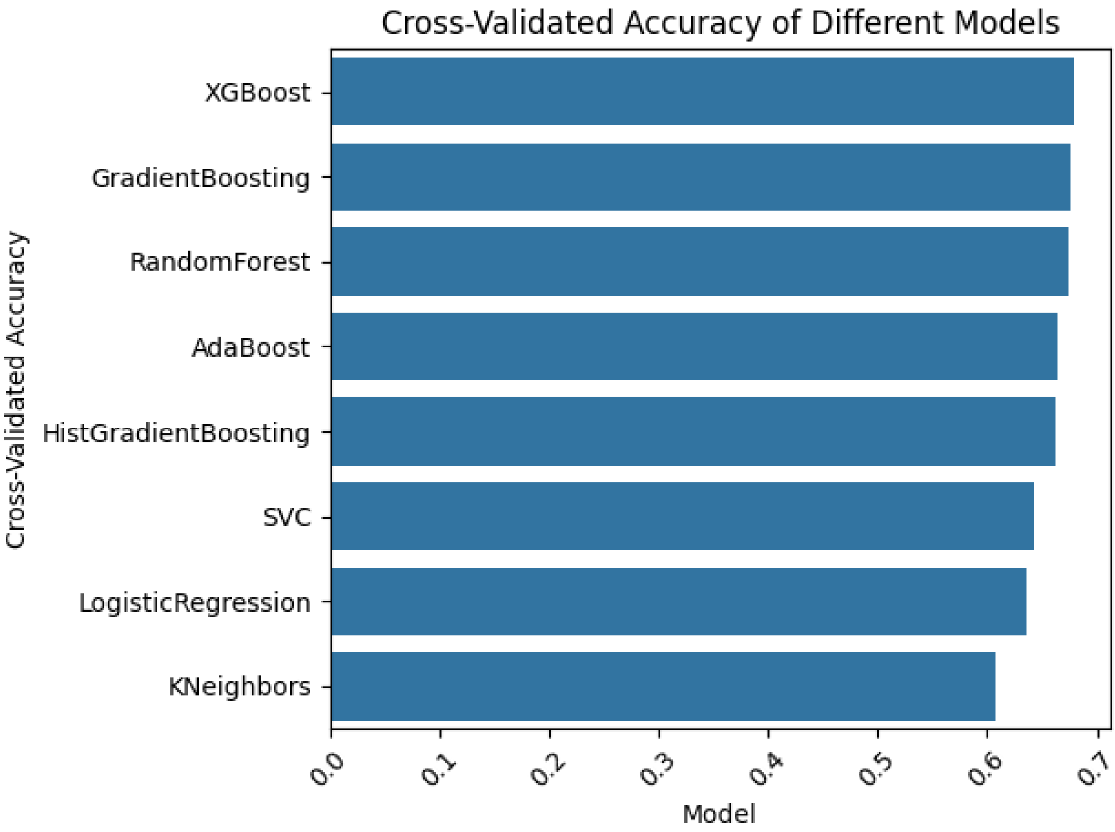


Figure 4: Evaluation of Different Models.

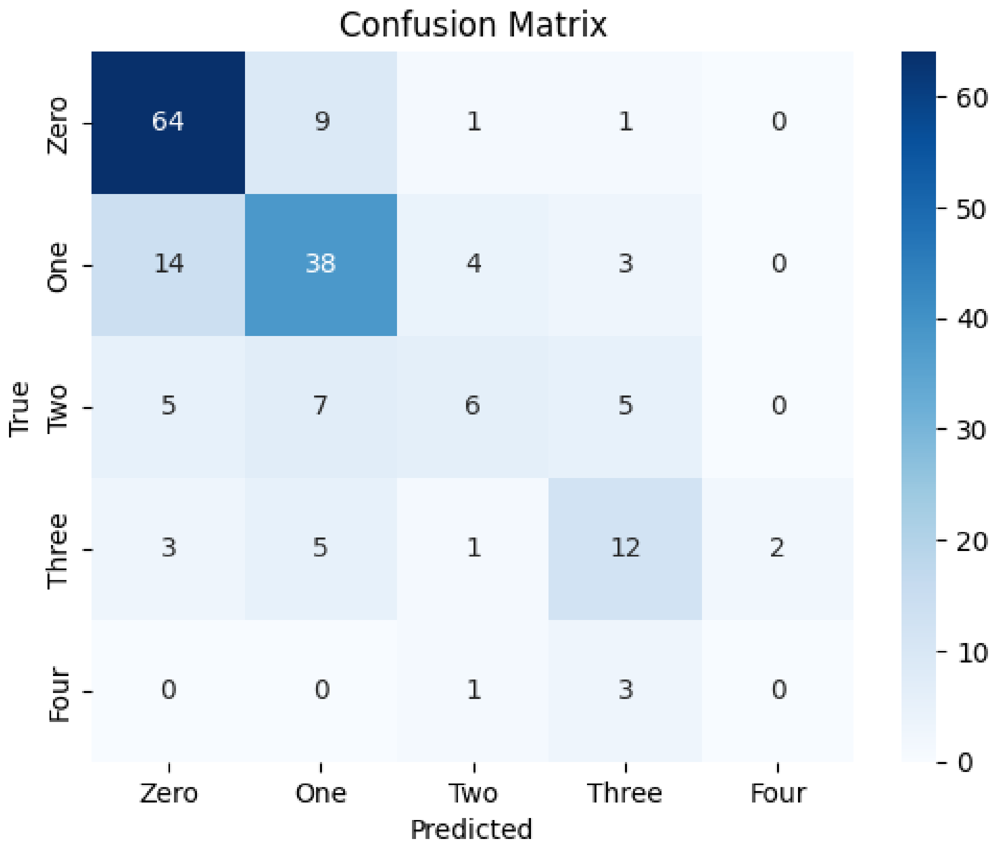


Figure 5: Confusion Matrix.

After the baseline assessment, a number of improvement strategies were used to maximise model performance with great success. Best among these was the combination of LightGBM and SMOTE, which tackled class imbalance directly,

best test accuracy (0.6739) and weighted F1 score (0.6630) in Figure 9, and a great ROC AUC value of 0.8941—a very well-balanced prediction across all heart disease severity levels. Additionally, the Stacking Ensemble stacking heterogeneous base learners like LightGBM, CatBoost, and HistGradientBoosting obtained top-tier performance with a test accuracy of 0.6685, a weighted F1 of 0.6558,

illustrated in Figure 10, and the best ROC AUC (0.8986). This indicates its improved class discrimination and generalisation ability. While Bayesian Optimisation with Optuna did not significantly outperform

GridSearchCV for HistGradientBoosting once again validated the model's insensitivity to tuning methods. On the other hand, simpler or more sensitive models like SVC, Logistic

Regression, along with shallow neural networks, remained behind, attesting to the efficacy and strength of ensemble and imbalance-handling techniques for formalised clinical datasets. These gains collectively advanced the modelling pipeline to a level of improved reliability, balance, and predictability.

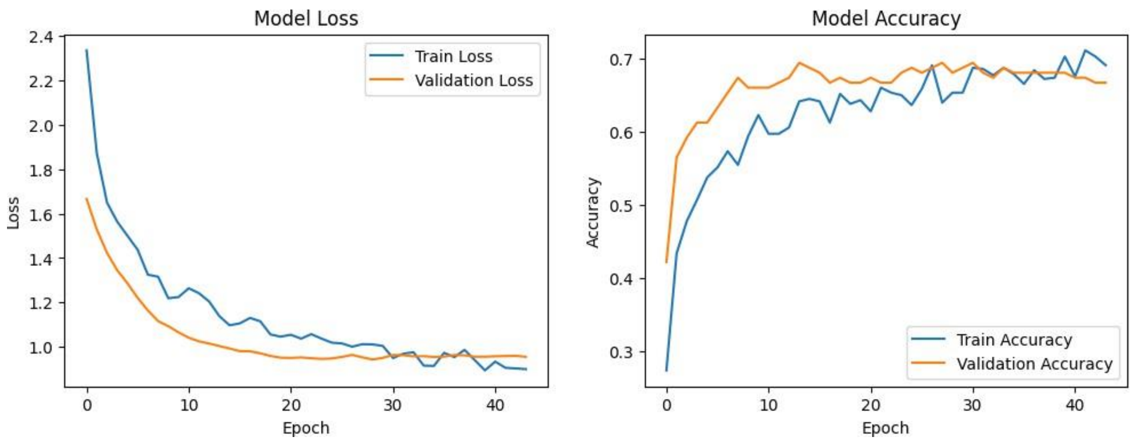


Figure 6: Training and Validation Loss every epoch.

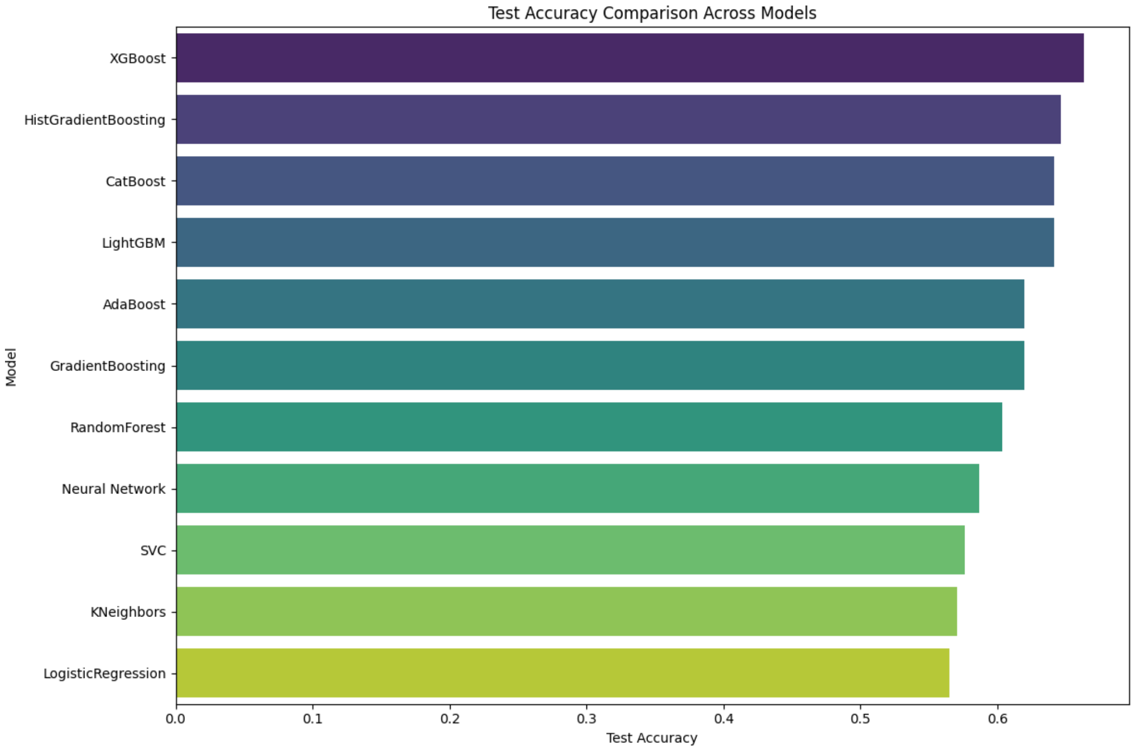


Figure 7: Accuracy Comparison Across Models.

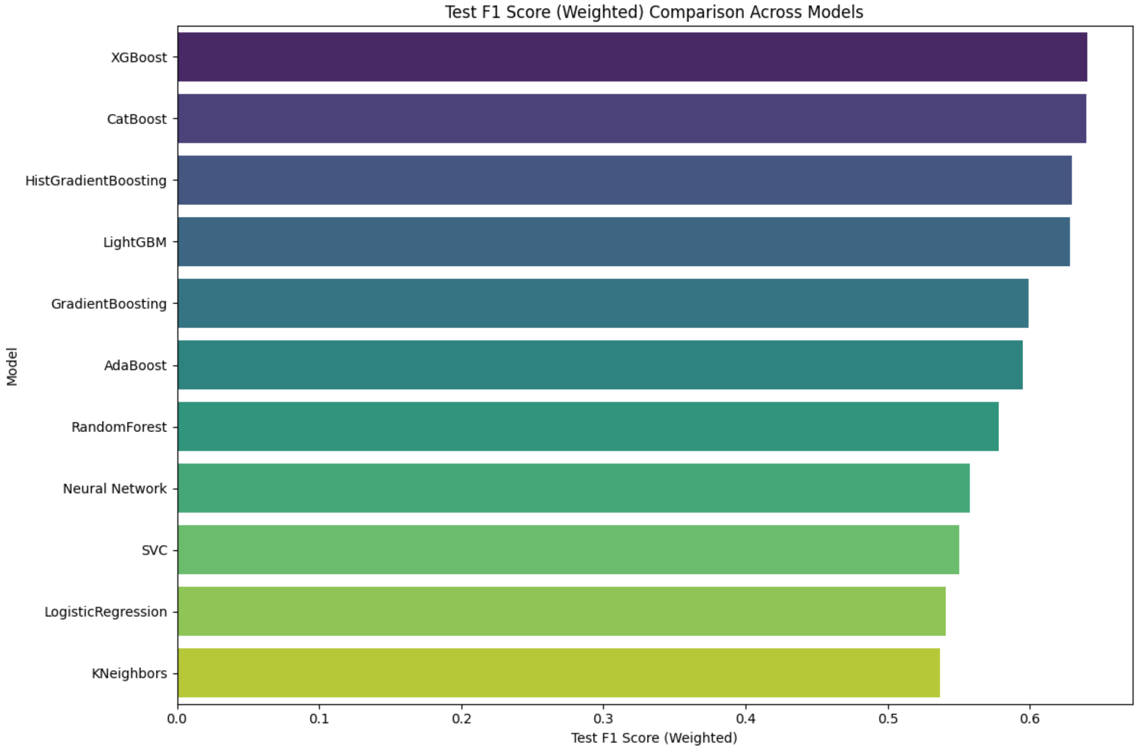


Figure 8: F1-score Comparison Across Models.

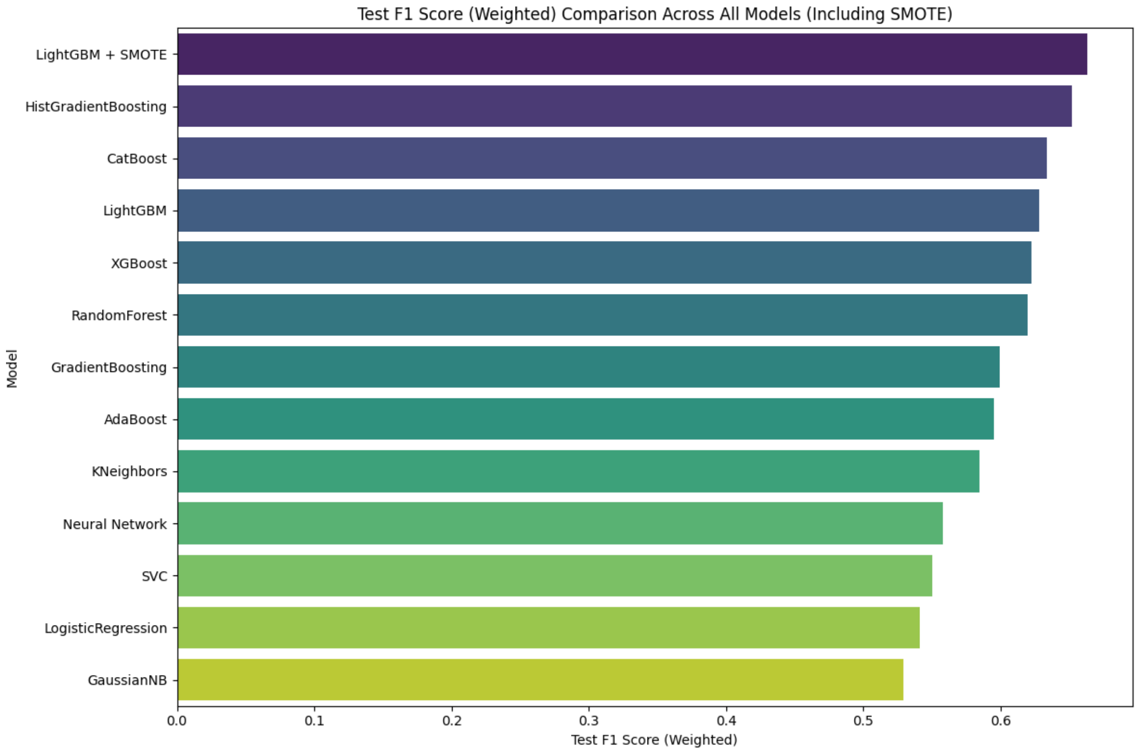


Figure 9: F1-score Comparison Across Models (Include SMOTE).

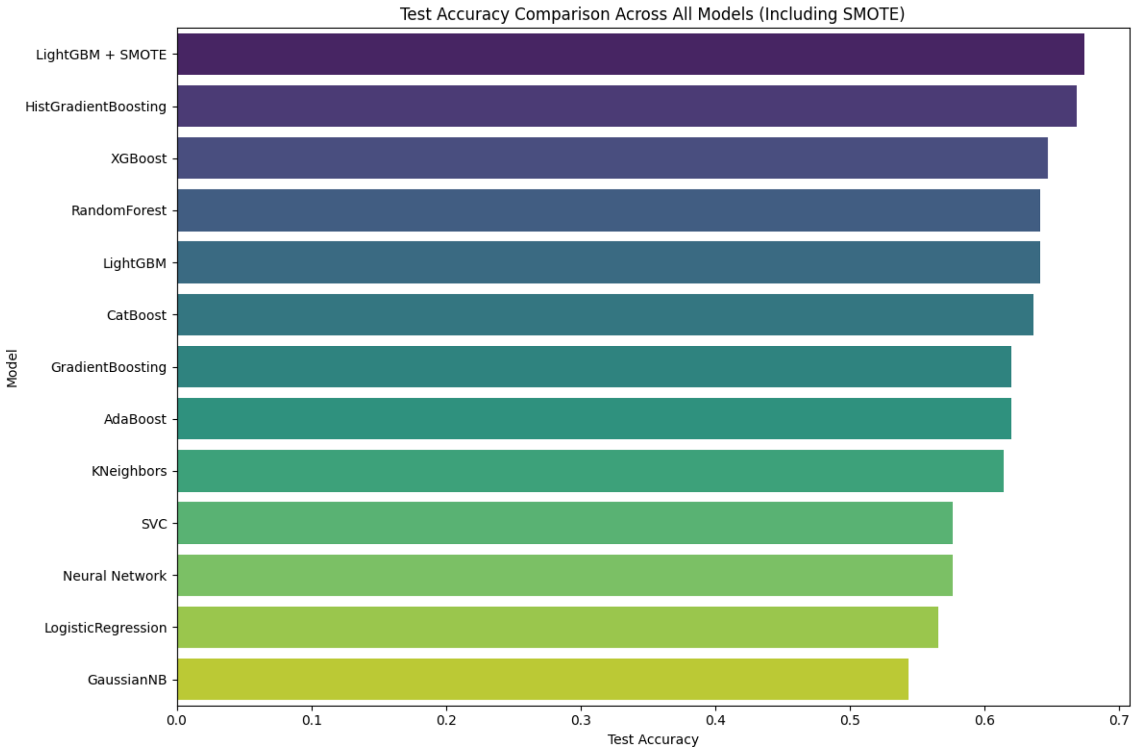


Figure 10: Test Accuracy Comparison Across Models (Include SMOTE).

Discussion

The present study results emphasise the ability of machine learning models in predicting the severity of heart disease, particularly with ensemble methods and solutions for handling data imbalance. The improvement in model performance of models like LightGBM with SMOTE stresses the importance of proper preprocessing as well as model selection in achieving high accuracy and stable predictions. Significant predictors like age and exercise-induced angina reinforce their clinical value in risk stratification and decision-making.

Conclusion

Cardiovascular disease (CVD) continues to remain among the most significant world health concerns, causing a considerable percentage of morbidity and mortality worldwide. Prevention of the CVD burden is of utmost importance before prediction and diagnosis, especially among young people, where non-modifiable and modifiable risk factors are of major significance. A review of literature indicates that while traditional statistical techniques have been effective, machine learning (ML) and deep learning techniques have provided greater precision, robustness, and adaptability in predictive modelling.

The integration of feature selection methods, data preprocessing, and class imbalance adjustment using techniques such as SMOTE has enhanced the performance of models such as logistic regression, random forest, XGBoost, and ensemble designs. Methods such as stacking ensembles and explainable AI (XAI) further add to model interpretability, generalisability, and clinical utility. Issues with dataset size, diversity, and biological interpretability persist, however, restricting the possible application of predictive models in routine clinical practice.

Future research should therefore concentrate on developing larger, representative databases, multimodal integration of health data (for example, IoT and biomarkers), and interpretable models that build clinician trust. Artificial intelligence can prove to be a game-changer for the early diagnosis, risk stratification, and management of cardiovascular disease, leading to enhanced patient outcomes and reduced global health burden, ending these challenges.

[1] S. R. Yeluri, H. K. Gara, and D. R. Vanamali, "Assessment of Knowledge with Regard to Cardiovascular Disease Risk Factors among College Students Using Heart Disease Fact Questionnaire Assessment of Knowledge with Regard to Cardiovascular Disease Risk Factors among College Students Using Heart Disease Fact Questionnaire," no. February, 2021, doi: 10.14260/jemds/2021/78.

[2] W. Degroat, H. Abdelhalim, K. Patel, D. Mendhe, S. Zeeshan, and Z. Ahmed, "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," *Sci. Rep.*, pp. 1–13, 2024, doi: 10.1038/s41598-023-50600-8.

[3] K. M. Shiwangi, J. K. Sandhu, and R. Sahu, "Effective Heart-Disease Prediction by Using Hybrid Machine Learning Technique," *Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2023*, pp. 1670–1675, 2023, doi: 10.1109/ICCPCT58313.2023.10245785.

[4] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthc. Anal.*, vol. 2, no. February, p. 100060, 2022, doi: 10.1016/j.health.2022.100060.

[5] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[6] A. S. Dina, A. B. Siddique, and D. Manivannan, "Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks," *IEEE Access*, vol. 10, no. August, pp. 96731–96747, 2022, doi: 10.1109/ACCESS.2022.3205337.

[7] A. Ishaq, S. Sadiq, M. Umer, and S. Ullah, "Improving the Prediction of Heart Failure Patients ' Survival Using SMOTE and Effective Data Mining Techniques," pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[8] D. Widhyanti and D. Juniati, "Heart disease prediction using machine learning techniques Heart disease prediction using machine learning techniques", doi: 10.1088/1757-899X/1022/1/012046.

[9] N. Varshney, M. E. M. Soudagar, and L. A. Al-keridis, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," no. April, pp. 1–9, 2023, doi: 10.3389/fmed.2023.1150933.

[10] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," vol. 2023, no. Cvd, 2023, doi: 10.1155/2023/1406060.

[11] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," vol. 2021, 2021, doi: 10.1155/2021/8387680.

[12] K. Vishnu, V. Reddy, I. Elamvazuthi, A. A. Aziz, and S. Paramasivam, "applied sciences Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," 2021.

[13] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," vol. 2022, 2022, doi: 10.1155/2022/7351061.

[14] J. A. Jevin, H. Jayant, R. Sanjay, V. Hemasai, and P. V Venkatasrinivas, "Heart Disease Identification Method Using Machine Learning Classification in," vol. 10, no. 3, 2023.

[15] G. N. Ahmad and S. H. Akhter, “Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection,” vol. 10, 2022, doi: 10.1109/ACCESS.2022.3153047.

[16] A. Tiwari, A. Chugh, and A. Sharma, “Ensemble Framework for Cardiovascular Disease Prediction”.

[17] F. K. Alarfaj, “XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques,” 2022.

[18] Available online: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data (accessed on 10 November 2023)."

[19] F. Almalki, A. A. Sheikh, I. Technology, and I. Technology, “OWE-CVD : A N O PTIMIZED W EIGHTED E NSEMBLE,” vol. 16, no. 3, pp. 29–45, 2025, doi: 10.5121/ijaia.2025.16303.