Journal of One Health Advances (JOHA)
Vol. 01, No.01, pp.01--13, March 2026

Predictive Modeling of Coronary Artery Disease: Data Preprocessing

Syed Mohammad Moiez Ur Rahman, Sujoy Mondol, Asjad Moiz Khan, Laxmi Kant Sagar, Preeti Dubey

Sharda University, Greater Noida, India

moiezr161@icloud.com, sujoy.mondol5r8@gmail.com, asjadk2021@gmail.com, laxmi.sagar1@sharda.ac.in, preetidubey.19dec@gmail.com

ABSTRACT

Coronary Artery Disease (CAD) is a global cardiovascular health problem with ever-growing rates, emphasising the need for convenient, reliable and efficient diagnostic methods. This study explores the use of ML based methods and algorithms for the early prediction of CAD using the Heart Disease dataset from the UCI Repository. The various challenges posed by the dataset, which include missing or noisy data and unbalanced sampling of the target variable, have been addressed in this study. A comprehensive machine learning pipeline was implemented wherein multiple data imputation methods were tested, the dataset was cleaned and further balanced using SMOTEENN. The balanced dataset was validated using a stratified split, and the dataset was then used to train nine ML models, including LR, SVM, Naïve Bayes, tree-based models and custom ensemble models. Hyperparameter tuning of the models was done using GridSearchCV. The custom Voting Ensemble model achieved the highest accuracy of 96.53% and an AUC of 0.98, followed by the custom Stacking Ensemble model with 95.95% accuracy, rest of the models achieved an accuracy greater than or equal to 93.18% indicating high preprocessed data quality. The results demonstrate the importance of high data quality and the effectiveness of ensemble models in capturing underlying patterns within patient data for clinical applications.

Keywords: *Coronary Artery Disease, Machine Learning, Predictive Modeling, Risk Diagnosis, CAD Prediction*

Introduction

Coronary Artery Disease (CAD) is among the leading global cardiovascular health problems today with growing rates. It occurs when the arteries supplying blood to the heart become narrow or blocked, leading to serious medical complications and even mortalities. As more recorded medical data becomes available everyday, machine learning has become a valuable and effective tool for assisting physicians in identifying CAD early on by identifying patterns in patient data that may otherwise go overlooked. But using machine learning for medical diagnosis comes with its own set of difficulties, such as class imbalances, noisy data, missing values, and variable feature importance.

The dataset, which was sourced from the UCI Machine Learning Repository [1], includes a wealth of patient data gathered from non-invasive tests, patient interviews, and medical examinations. Along with inconsistent data and a notably high number of missing values, the dataset presents serious difficulties and necessitates sufficient data preprocessing. A thorough data imputation procedure was used to address the missing data. To determine which type of imputation to analyze was best, the missing data was first imputed and compared using three distinct techniques: Simple Imputation, kNN Imputation, and MICE Imputation. To maintain data integrity, outliers and noise were also addressed using statistical analysis and domain expertise. In order to successfully address the class imbalance problem and eliminate noise, SMOTEENN was used to balance the unbalanced samples of the target variable that were present in the imputed datasets. To ensure distribution consistency across splits, stratified 5-fold cross validation was then used to validate the balanced dataset.

*Corresponding author: Syed Mohammad Moiez Ur Rahman, Sharda University, Greater Noida, India (moiezr161@icloud.com)

In order to fully assess predictive performance, nine machine learning models, including LR, SVM, Naive Bayes, tree-based classifiers, and three custom ensemble models with different base learners were trained using the preprocessed data. In order to fine-tune the obtained accuracies, the output of each model was followed by the proper hyperparameter tuning using the Grid Search CV technique. A greater emphasis on data preprocessing can significantly increase the accuracy of ML models, based on the findings described in this study. The models can better handle irregularities in real-world data by experimenting with different imputation strategies and using robust resampling techniques like SMOTEENN.

Literature Survey

In the recent comparative study done by Beri et al. (2024) and his colleagues worked on three classification algorithms which are aggressive for the heart disease prediction as they showed a decent amount of growth in the accuracy part with logistic regression having the highest accuracy of 82.8% with random forest as 78.1% and decision trees at 70.3%, they mainly worked on achieving and solidifying the growth potential of the machine learning in cardiovascular diagnosis [2]. Rather than just studying on the machine learning algorithm and their features, there's Rana et al. (2025) and colleagues who addressed a critical challenge identifying the cardiovascular disease when traditional tests like electrocardiograms (ECGs) looked normal even when the patient was suffering from heart problems. The researchers greatly emphasized on advanced imaging technologies like specialized CT scans and MRI techniques to enhance the diagnostic precision in complex cases, the techniques include SVM, DT, ANN and Deep Learning models like DNNs and CNNs along with tools as Computational Magnetic Resonance and Cardiac Computed Tomography (CMT) [3].

Jha et al. (2025) used the dataset called Cleveland Heart Disease Dataset to test up against the fact how well machine learning could help in predicting heart disease, their study thoroughly covered the data preprocessing steps with outlier removal as first, then categorical encoding enforcing numbers so models can process with missing value imputation with statistical methods and feature scaling. Their main models included SVM, DT, RF and ANN trained and evaluated on 80:20 train-test split with ANN presenting the highest accuracy of 86%, 86% precision, 84% recall and an F1-score of 83% [4]. Ezekwueme et al. (2025) worked on the fact that how deep learning models and natural language processing can help doctors diagnose CAD using non-invasive methods. Their team kept a check on the data privacy also including the expensive computing requirements and unequal access to the particular set of technology [5].

Sun et al. (2025) introduced "radiomics" - a technique that uses computers for extracting subtle mathematical patterns and features from medical images which are nearly invisible to the human eye, the process of finding hidden fingerprints in image data, the problem they are faced up with was Acute Ischemic Stroke (AIS) occurring when blood flow to part of the brain which is blocked. The challenge was pressed as the damaged brain tissue doesn't show up on standard CT scans, even though the stroke has actually occurred within 24 hours creating a threatening diagnostic gap where doctors might miss early strokes. Their research team studied on 1,122 patients with confirmed strokes who initially showed "negative" CT scans but later MRI scans proved they actually had strokes. Their research included 8 hospital's patient data (5 for training, 3 for validation), in image processing part, they took technical steps like head alignment and image co-registration to standardize the scan with identifying 44 key radiomic features and were evaluated on SVM, RF, LR and NN resulting in AUC scores of 0.80 which further enhanced early AIS diagnosis from standard NCCT scans [6].

Vitorino (2025) presents a comprehensive study of proteomics approaches where different proteins which appear in blood and urine when someone has coronary artery disease and these proteins break down as the biomarkers stating as the early warning signs of disease. Two main approaches are

compared with first being urine analysis which is completely non-invasive and can be repeated frequently without discomfort but posing with a challenge as it contains lower concentration of protein, then the next is blood analysis which is more invasive but give out the richer comprehensive protein profiles useful for urgent diagnosis. The researchers proposed on combining protein data from multiple sources with AI for creating personalized treatment plans for each patient. The limitations of this research is that it needs more standardized collection methods, testing across diverse populations to combine different types of data. In order to fully apply proteomics for CAD, their study's conclusion emphasizes the necessity of uniform collecting techniques, wider biomarker validation across populations, and enhanced data integration [7]. Alsharqi and Edelman (2025) review the transformative role of AI in cardiovascular imaging for interpreting complex medical images from multiple sources simultaneously, while deeply emphasizing on the valuableness for complicated conditions like heart failure and irregular heart rhythms (atrial fibrillation). The role of AI in procedural planning has always helped the doctors plan heart procedures more effectively and the real-time guidance which optimizes image quality during procedures like cardiac catheterization. The researchers acknowledge the fact that AI could helped to ease up the procedural workflow in the health industry but it should work safely and transparently as per in clinical settings [8].

Tolu-Akinnawo et al. (2025) introduces AI across all heart imaging methods which gives a broad overview of how AI is being used across all major types of heart imaging - Echocardiography, MRI, CT, Nuclear Imaging. The use of AI in health industry can specifically highlight the ML and DL models so they can detect “subtle anomalies” - very small or early signs which are often overlooked as AI can great with the image enhanced quality to clear up the blurry or noisy images with appropriate model training, but it also comes with some highly addressing issues like regulatory oversight, data quality issues, standardization problems. The researchers argue that AI’s success is directly proportional to the collaboration between doctors, technology companies, and regulatory agencies working together [9]. Bednarek et al. (2025) introduced the need of AI for personalized CAD treatment majorly focusing on the specific type of AI which is good at analyzing the images - Convolutional Neural Networks (CNNs), they are designed in such a way so that it can interpret data automatically from the CCTA (Coronary Computed Tomography Angiography) which included the detailed CT scans of heart arteries and fetching the data from the OCT (Optical Coherence Tomography) which states the high-resolution imaging of blood vessel walls. This study emphasizes the need to integrate AI with hospital management and early diagnosis of patients to use the resources more effectively, this way clinicians can improve both clinical outcomes and a lack of large-scale randomized trials. The clinical applications as per within hospital environment will become far better as it will have real-time data to train with and also provide several key benefits which include risk assessment of the patient, prediction modeling of how diseases will progress, treatment personalization for choosing the best treatment for each individual patient [10].

Singh et al. (2025) presented Adaptive Gated Spatial CNN as per ultrasound CAD detection which includes CNN, the spatial part which phrases how the network pays special attention to where things are located in the particular image and how they relate spatially to each other and it uses “gates” mainly the mathematical filters that decide which information must be there for the enhancement and the final point where the system can adjust its approach based on what it learns from each image. This technique is non-invasive, cost-effective and provides real time immediate results. The research also concluded with 96.45% accuracy with correctly identifying the CAD presence in 95 out of 100 cases, 90.45% sensitivity, 94.36% specificity, 94.56% ROC score. This represents one of the major advancement as ultrasound has traditionally been less reliable for CAD detection compared to other imaging methods. The high sensitivity helps minimizing the missed diagnoses which is particularly important as clinically missing CAD can be life-threatening [11]. Rajeev and Natarajan (2025) conducted CT Angiography-

Based detection for CAD in their recent study, the main agenda for this research was visual representations that show which parts of the image needs to be focused on with the term called “heatmaps”.The technical innovation which worked around this research is involving a ConvMixer which is a newer type of neural network architecture that combines convolution with mixing (information integration) in a much better and efficient way than the traditional CNNs. The research’s main two image processing points are median filtering which helps in removing noise and artifacts from images and the next one is morphological operations including mathematical techniques which enhance structural features found in images. The research used 5,959 CT angiography images to introduce robustness to the model with 96.30% accuracy, 94.39% sensitivity, 99.16% specificity, all helping in the clinical acceptance for more doctors to trust and understand AI recommendations [12].

Ried et al. (2025) explored how combining machine learning with CT-derived Fractional Flow Reserve (CT-FFR) could make heart disease diagnosis more accurate and less invasive. They studied 421 patients who had both coronary computed tomography angiography (CCTA) scans and follow-up tests like cardiac MRI and invasive angiography. Their results showed that using CT-FFR alongside traditional CCTA scans significantly improved the ability to spot ischemia, a condition where the heart does not get enough blood. The combined method reached an AUC of 0.87, outperforming CCTA alone, which had an AUC of 0.83 ($p < 0.0001$). This suggests that CT-FFR could help doctors diagnose heart problems more precisely without having to rely on invasive tests, making the process safer and more comfortable for patients [13]. Pezel et al. (2025) introduced a machine learning model designed to predict major adverse cardiovascular events (MACE) in patients newly diagnosed with obstructive coronary artery disease (CAD), aiming to enhance long-term risk stratification. The study consisted of 2,038 patients with a median follow-up of seven years, the researchers integrated clinical, ECG, coronary CT angiography (CCTA), and stress cardiac MRI data using an XGBoost algorithm with LASSO-based feature selection. The model demonstrated strong predictive capability, achieving an AUC of 0.86, notably outperforming conventional risk scores such as the European Society of Cardiology score (AUC 0.55), QRISK3 (0.60), and the Framingham Risk Score (0.50), as well as models based solely on CCTA (0.76) or MRI (0.83). It also held robust performance in external validations (AUCs of 0.84 and 0.92). These findings underscore the potential of advanced, multimodal machine learning approaches to deliver more personalized and accurate prognostic insights for patients with obstructive CAD [14].

Mirjalili et al. (2025) demonstrated a practical approach to estimating coronary artery disease (CAD) severity by combining routine ECG readings with common risk factors such as hypertension. Their study found that certain ECG changes were closely associated with higher SYNTAX and Gensini scores, which reflect the extent of arterial blockages. This finding suggests that a simple and widely accessible ECG could help clinicians identify patients who need more urgent or detailed testing, offering a valuable tool especially in settings where resources are limited [15].

Recent research in heart disease prediction has largely emphasized advanced and diverse data collection and diagnostic techniques, often overlooking existing datasets like the Cleveland Heart Disease dataset, which contains inconsistencies and imperfections that mirror real-world clinical data. This study addresses that gap by focusing on rigorous preprocessing to enhance the quality of the Cleveland dataset. The findings demonstrate that, with proper refinement, even imperfect and inconsistent data can yield clinically reliable and accurate diagnostic results.

Research Methodology

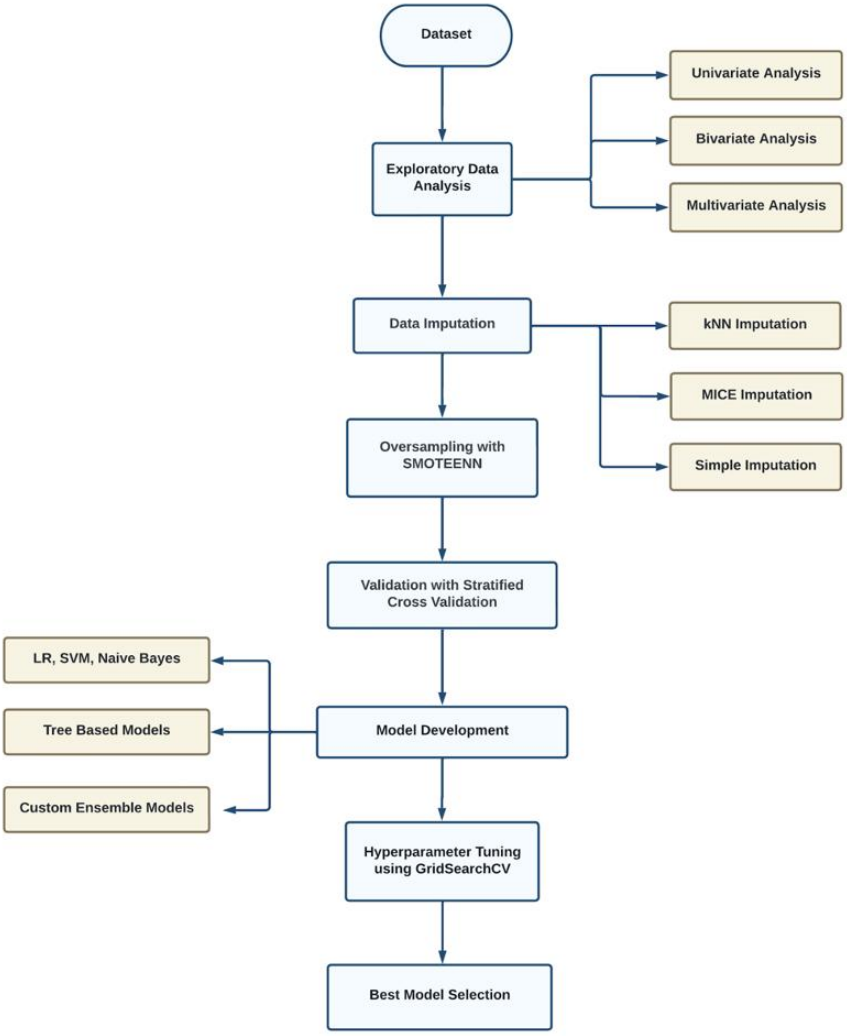

Figure 1. Methodology Flowchart

The research methodology, shown in Fig. 1 starts by collecting the dataset, in this case acquired from the UCI Machine Learning Repository. Exploratory Data Analysis is then performed on the dataset to better understand it, these including visualizing missing values, univariate analysis of each attribute, bivariate analysis of an attribute to understand how does it change with respect to the target variable and multivariate analysis. A comprehensive data imputation strategy was then employed with various imputation methods such as Simple Imputation, kNN Imputation and MICE Imputation. The imputed datasets were then investigated for class imbalances in the target variable, which were then catered using the SMOTEENN (SMOTE+ENN) method. Validation was perfomed on the balanced dataset using a stratified split of 5-fold cross validation. The preprocessed dataset was then used to train nine ML models which include: LR, SVM, Naïve Bayes, tree-based models and custom-ensemble models. The models were then hyperparameter tuned to fine tune their output with GridSearchCV, after which the results of the models were recorded and visualized to select the best model.

The dataset, obtained from the UCI Machine Learning Repository is often used to predict coronary artery disease due its mature and reliable patient information. The dataset has 76 features; however, most studies only use 14 due to extensive missing values and data redundancy across various features.

The dataset includes details like blood pressure, cholesterol levels, ECG results, and exercise test results, all gathered through non-invasive tests.

For the purpose of this study, the subset of attributes that are commonly used in existing literature were selected, allowing for the increased focus on features with established clinical relevance. By narrowing the scope to these well studied variables, the study aims to conduct a more in-depth analysis while also enhancing model reliability through improved data quality and integrity. Addressing these missing values is a critical step, which gets more crucial and challenging due to the presence of both numerical and categorical variables. The target variable is ‘num’ which contains the values: 0, 1, 2 and 3. In the dataset, ‘0’ indicates no CAD, ‘1’ indicates mild CAD, ‘2’ indicates severe CAD and ‘4’ indicates critical CAD. The target column was processed to contain only binary values, including only, ‘True’ which was categorized for patients having the value 2,3 or 4 as they would be under the risk of CAD, and ‘False’ for patients having the value 0 as it depicts no risk of CAD.Exploratory data analysis was performed to better understand the various relationships in the dataset, which included the univariate, bivariate and multivariate analysis. The univariate analysis provides insights into the distribution of individual features, identifying the presence of outliers, especially in the numerical attributes, the bivariate analysis of the dataset explored relationships between features and their association with the target variable.

Data imputation was a challenge due to the high amount of missing values and the presence of both numerical and categorical features. Three separate types of data imputation methods were applied to three separate copies of the dataset, as shown in Fig. 1, to test their efficiency in increasing the quality of the dataset and successfully maintaining the underlying pattern of patient data. The first copy was treated using the simple imputation method, where the median of a feature column was used to impute the numerical columns with missing values. The categorical columns contained a lesser percentage of missing values, resulting in them being imputed using the mode of the feature. The second copy was imputed using the kNN imputer, it fills missing values based on the average (for numerical) and the mode (for categorical) of the nearest neighbors, effectively preserving local patterns and similarities within the data. The third copy used MICE (Multiple Imputation by Chained Equations), it models each feature with missing values as a function (eq. 1) of the others in an iterative process, capturing complex interdependence and producing more statistically robust imputations for both numerical and categorical features.

$$\hat{Y}_j^{(t)} = E\left[X_j \mid X_{-j}^{(t)}\right] + \varepsilon_j \qquad (1)$$

Where:

$\hat{Y}_j^{(t)}$ is imputed value for feature X_j at iteration t

$X_{-j}^{(t)}$ is all other features except X_j using their most recent imputations

$E\left[X_j \mid X_{-j}^{(t)}\right]$ is the expected value of X_j given the other variables (typically from a regression model)

ε_j is the noise term to account for uncertainty in imputation

Several instances in the feature columns were dropped as they either bypassed null tests but were not meaningful or contained invalid data. For instance, rows with a value of '0' in the ‘trestbps’ column, which represents resting blood pressure, were removed since a value of '0' for resting blood pressure is not physiologically possible and indicates missing or erroneous data. Similarly, the ‘exang’ column, which represents exercise-induced angina and is expected to contain binary values ('TRUE' or 'FALSE'), included several instances with 'NaN' values, which were identified and dropped. The ‘restecg’ column,

which records the resting electrocardiogram (ECG) results, also contained instances with 'nan' values, leading to the removal of those rows to maintain the integrity of the dataset. Duplicate rows can artificially inflate model performance and bias the training process. By retaining only unique records, the dataset is both accurate and representative, ultimately enhancing the reliability of the models.

To address the class imbalance, the SMOTEENN algorithm was used, SMOTEENN is a hybrid resampling technique combining both oversampling and undersampling strategies. It integrates SMOTE (Synthetic Minority Over-sampling Technique), a technique that synthetically generates new instances of the minority class by interpolating between existing minority class samples, and Edited Nearest Neighbours (ENN), an undersampling method that removes ambiguous or misclassified examples from the majority class based on their nearest neighbors (eq. 2). The dual mechanism balances the class distribution by enhancing minority class representation and also denoises the dataset by eliminating noisy majority class instances that could hinder the model's learning process, this is particularly important as highly complex imputation methods used earlier, such as kNN and MICE can introduce noise to smaller dataset.

$$\tilde{x}_i = x_i + \lambda \cdot (x_{ni} - x_i) \qquad (2)$$

Where:

x_i is the original minority sample,

x_{ni} is one of its k-nearest neighbors,

$\lambda \in [0, 1]$ is a random scalar,

$\tilde{x}_i$ is the generated synthetic sample.

To evaluate model reliability, 5-fold stratified cross-validation was employed, ensuring class distribution remained consistent across splits. SMOTE was applied only to the training folds to generate synthetic minority samples and avoid data leakage. A Random Forest classifier was then trained and tested on each fold. Across all three imputed datasets, the model showed strong and consistent results, achieving an average accuracy of 90% and high scores across other metrics, highlighting the effectiveness of the resampling and validation approach in handling class imbalance.The models were trained on the algorithms: LR, RF, XGBoost, SVM, Naive Bayes and three custom ensemble models. The first ensemble model, the voting ensemble had LR, RF and XGBoost as the base learner and followed a soft voting approach, the second ensemble, stacking ensemble utilized the same base learners, where each one was treated as a feature for the final meta model, the third ensemble, bagging ensemble had the DecisionTreeClassifier as the base learner which was used with the BaggingClassifier where multiple independent models were then trained on the various different subsets of data and their result was aggregated, significantly aiding in avoiding overfitting and variance.Following model training on datasets processed with different imputation techniques, the models trained on the dataset imputed using Simple Imputation consistently outperformed the more complex methods, suggesting that the latter may have introduced noise into the dataset, potentially distorting underlying data patterns and reducing model performance. As a result, Simple Imputation was selected for all subsequent analyses.

Hyperparameter tuning was done using GridSearchCV. For hyperparameter tuning, the parameters for each model was selected as follows: For gradient boosting models such as XGBoost and LightGBM,

similar parameters were adjusted. varying the number of boosting rounds (n_estimators) and controlling the learning rate (learning_rate) with values such as 0.01 and 0.1, helping the model learn more gradually and reducing the risk of overfitting. The max_depth parameter was tuned to control the complexity of the trees, and for LightGBM specifically, the number of leaves (num_leaves) was altered to capture more or fewer splits per tree, directly affecting model complexity and learning capacity. Logistic Regression and Support Vector Machine (SVM) models were optimized by adjusting the regularization parameter C, responsible for balancing the trade-off between achieving low training error and maintaining a simpler model, the values 0.1, 1, and 10 were explored. The penalty was fixed to 'l2' for regularization, and solvers like 'lbfgs' and 'liblinear' were compared based on convergence speed and dataset suitability. For SVM in particular, the kernel type ('linear' or 'rbf') and the gamma parameter were tuned, responsible for determining how far the influence of a single training example reaches. Lastly, for Naïve Bayes, the var_smoothing parameter was adjusted across a range of very small values (from 1e-09 to 1e-06) to ensure stability during calculations and reduce sensitivity to numerical precision errors. All hyperparameters were systematically explored using GridSearchCV to identify the optimal configurations that maximize model performance.

Results

The dataset, obtained from the UCI Machine Learning Repository [1], contains 76 attributes, out of which only 14 were used for this study. Out of the 14 selected attributes, 10 of them had missing values. The dataset contains missing values, redundant data and other imperfections, mimicking real life data where all the data is not present in consistent and perfect form.

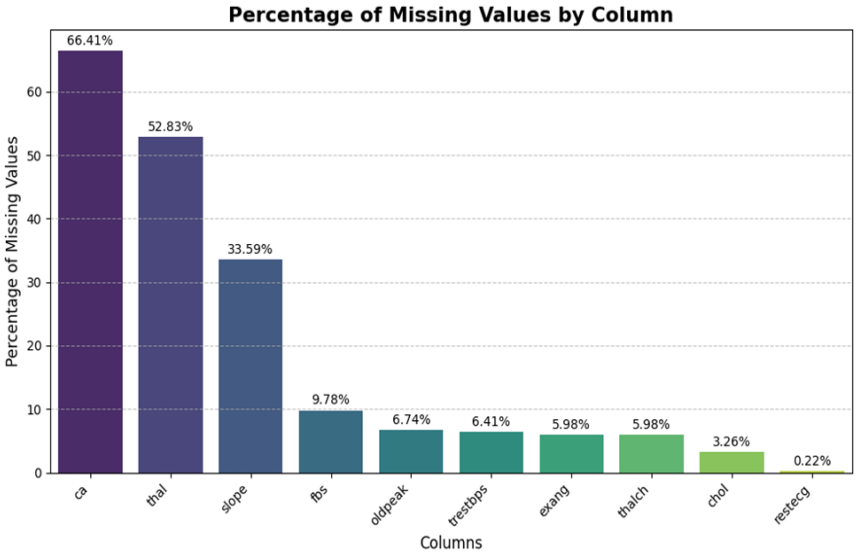

Figure 2. Percentage of Missing Values by Column

Fig. 2 displays that the dataset contains missing values across several features, with some attributes exhibiting significant proportions of missing data. Notably, features such as ‘ca’ and ‘thal’ have missing values of 66.41% and 52.83%, respectively, while others like ‘slope’ and ‘fbs’ show smaller missing percentages. The missing values were addressed using the Simple Imputation, kNN Imputation and MICE Imputation techniques.

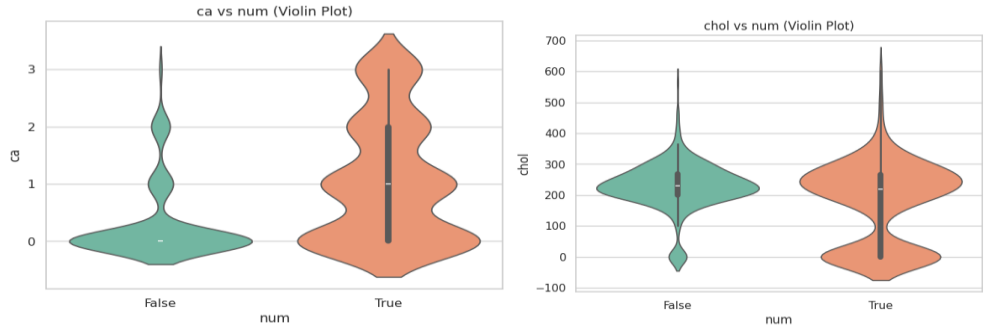

Figure 3. Violin plot (Bivariate Analysis)

As shown in Fig. 3, the violin plots show how two features, 'ca' (number of major vessels) and 'chol' (cholesterol level) vary between patients with and without heart disease. The first plot shows that most patients without heart disease have 'ca' values close to zero, while those with heart disease tend to have higher values, suggesting a possible link. In the second plot, cholesterol levels appear to be spread out across both groups, but there's a slightly wider range and higher concentration of values in patients with heart disease, which might indicate a weak association. The visual and statistical insights not only guided the selection of robust strategies for handling outliers during simple imputation (such as using the median) but also contributed to a deeper understanding of how specific attributes vary with respect to the target class, ultimately improving the overall quality of the data preprocessing pipeline.

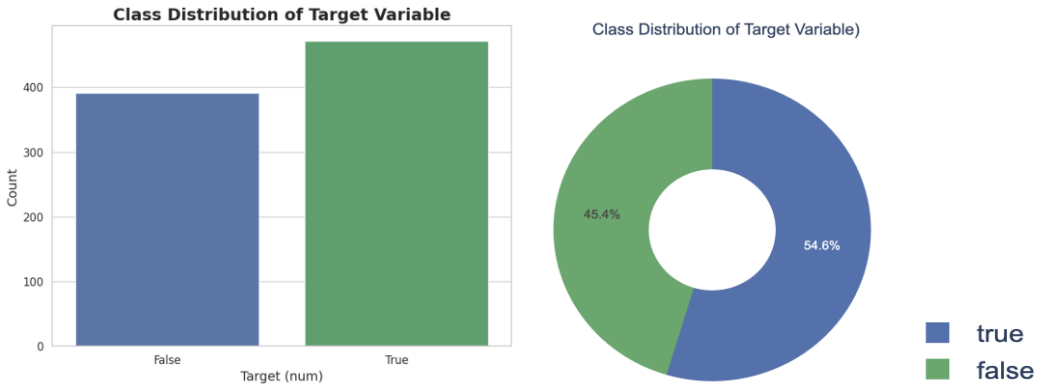

Figure 4. Class Balance in the Target Variable

The imputed dataset had a class imbalance of the target variable. Fig.4 shows the majority class comprised of the instances of the individuals having a risk of developing CAD (54.6%) and the minority class comprised of the instances of individuals not at a risk of CAD (45.4%) in the target variable. The balancing of the target variable samples was done using SMOTEENN. The sample imbalance may seem negligible, however, considering the size of the dataset, it could be deemed significant. The class imbalance, with a higher number of individuals at risk for CAD, could lead to a bias in the model towards the instances of individuals having CAD, potentially reducing its ability to accurately identify those not at risk.

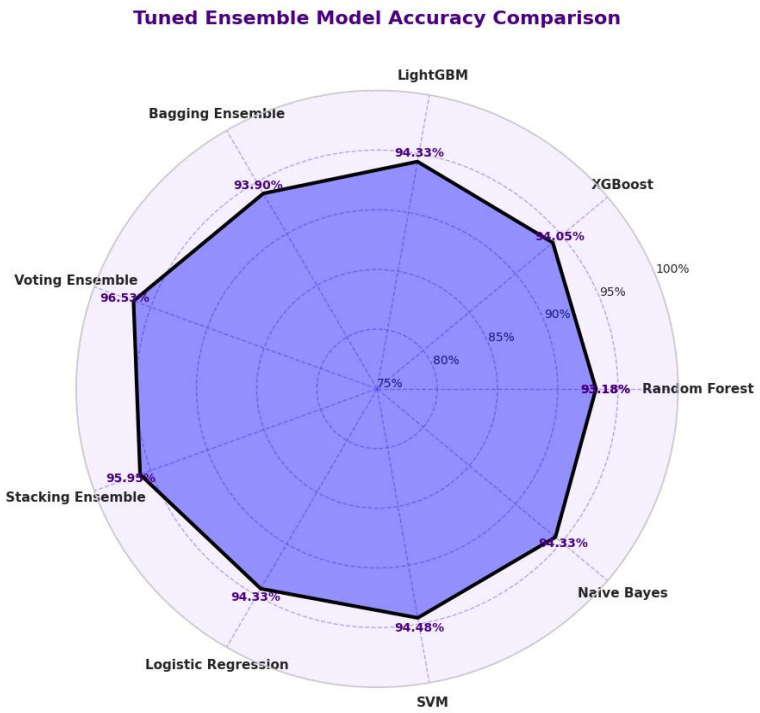

Figure 5. Analysis of the Tuned Model Accuracies

The trained models were subjected to hyperparameter tuning using GridSearchCV, which systematically explored a predefined grid of parameter values. The results of the tuned models are illustrated in Fig. 5, wherein, the custom Voting Ensemble Model emerged as the top performer achieving the highest accuracy of 96.53%. The custom Stacking Ensemble Model followed closely, with an impressive accuracy of 95.95%, reflecting the high performance of the custom ensemble models. All remaining models achieved competitive accuracy levels, with each model scoring greater than or equal to 93.18%, highlighting the effectiveness of the ML models when trained upon high quality data. The results also highlight the effectiveness of ensemble-based models in learning and capturing complex patterns among patient data.

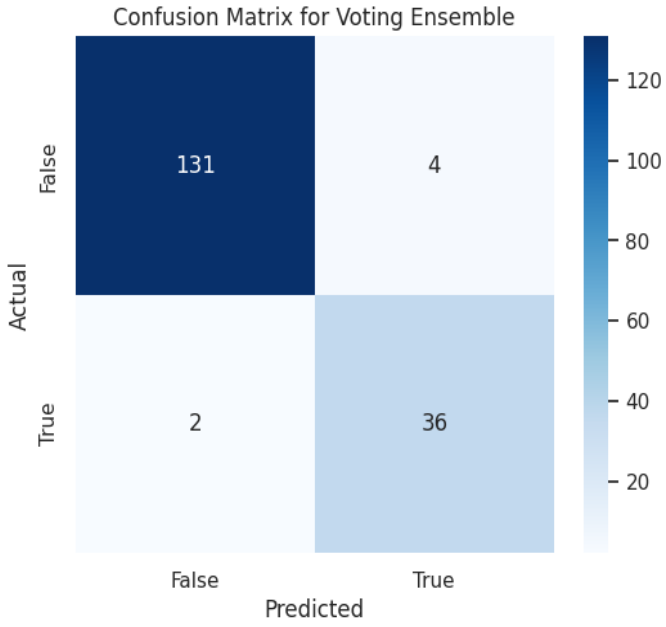

Figure 6. Confusion Matrix

Fig. 6 shows that the Voting Ensemble model had a recall score of 0.9653, the high recall score indicates that the model successfully captures most of the actual positive cases, reducing the risk of missed diagnoses. By effectively capturing most true positive instances, the model enhances diagnostic reliability, making it a valuable tool for supporting early and accurate identification of CAD in medical applications This is particularly significant as a higher recall ensures fewer false negatives, vital for identifying patients with CAD who may otherwise go undetected.

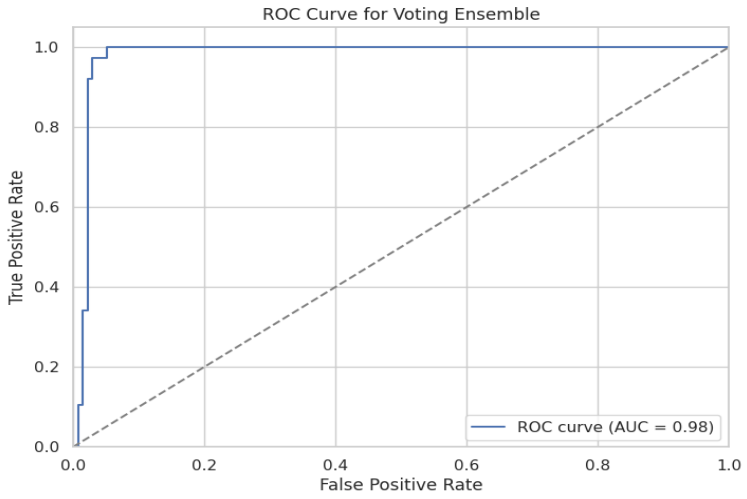

Figure 7. ROC Curve of the Voting Ensemble Model

A clear picture of the Voting Ensemble Model's performance in class distinction is provided by the ROC in Figure 7. The model rarely incorrectly determines negatives as positives while accurately identifying positives, as indicated by the steep curve towards the top-left corner, which also shows a low false positive rate. The model has a 98% chance of accurately differentiating between a randomly selected positive case and a negative one, according to it's AUC score of 0.98. The model's strong performance and possible dependability are reflected in its high AUC score.

Conclusions

The study aims to enhance the use of ML algorithms and pipelines in real world clinical settings. An imperfect and inconsistent dataset is explored in this study, mimicking real life data which often contains imperfections, irregularities and inconsistencies. Through rigorous preprocessing and the fine tuning of ML models, the study was able to achieve an accuracy of greater than or equal to 93.18% on all the trained algorithms, portraying how real world data, with its imperfections can also be used to train ML and achieve high diagnostic accuracies.

The integration of the ML models in the real world clinical settings, considering the current stage and reliability, could be in the form of a risk diagnosis. By classifying patient with non-invasive methods, it can motivate more people to get tested, promoting public health while also expanding the data it stores, further improving model generalization in real time. It can also help better utilize the resources of the medical staff by diagnosing the severity of CAD in each patient, ensuring efficient resource utilization of the clinicians and directing medical attention to the cases which need it the most.

The models in the study showed remarkable accuracy in spite of the drawbacks of having fewer instances and inconsistent initial data. In terms of classification ability and recall, which are crucial in the risk diagnosis of CAD and where reducing false negatives can have a direct effect on patient outcomes, the Voting Ensemble model in particular performed the best across a number of metrics. All models performed competitively, suggesting that the improved dataset quality significantly contributed

to model effectiveness. These results underscore the idea that strong preprocessing pipelines can compensate for dataset limitations and lead to clinically meaningful insights.

While the results of the study are promising, they can be improved further. The use of a relatively small dataset may not fully capture the variability of real-world CAD cases, potentially hindering the model's generalizability. Future research should focus on incorporating extensive and diverse amounts of real patient data to eliminate issues like missing values and improve the generalizability of the models. Integrating such models into clinical decision support systems could significantly assist in early detection and intervention of CAD, potentially motivating more people at risk to get tested, due to the non-invasive and passive nature, additionally, also helping prioritize medical attention towards cases based on the severity, optimizing medical resources and promoting patient diagnosis.

References

[1] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

[2] Beri, M., Gill, K. S., & Sharma, N. (2024, October). Advanced Machine Learning Techniques for Early Prediction of Heart Disease: A Comprehensive Analysis. In 2024 4th International Conference on Sustainable Expert Systems (ICSES) (pp. 1219-1223). IEEE.

[3] Rana, N., Sharma, K., & Sharma, A. (2025). Diagnostic Strategies Using AI and ML in Cardiovascular Diseases: Challenges and Future Perspectives. In Deep Learning and Computer Vision: Models and Biomedical Applications: Volume 1 (pp. 135-165). Singapore: Springer Nature Singapore.

[4] Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. J Cardiobiol, 9(1), 1.

[5] Ezekwueme, F., Tolu-Akinnawo, O., Smith, Z., & Ogunniyi, K. E. (2025). Non-invasive Assessment of Coronary Artery Disease: The Role of AI in the Current Status and Future Directions. Cureus, 17(2).

[6] Sun, K., Shi, R., Yu, X., Wang, Y., Zhang, W., Yang, X., ... & Wang, X. (2025). Noninvasive imaging biomarker reveals invisible microscopic variation in acute ischaemic stroke (≤ 24 h): a multicentre retrospective study. Scientific Reports, 15(1), 3743.

[7] Vitorino, R. (2025). Minimally Invasive Versus Invasive Proteomics: Urine and Blood Biomarkers in Coronary Artery Disease. PROTEOMICS–Clinical Applications, 19(1), e202400062.

[8] Alsharqi, M., & Edelman, E. R. (2025). Artificial Intelligence in Cardiovascular Imaging and Interventional Cardiology: Emerging Trends and Clinical Implications. Journal of the Society for Cardiovascular Angiography & Interventions, 4(3), 102558.

[9] Tolu-Akinnawo, O. Z., Ezekwueme, F., Omolayo, O., Batheja, S., & Awoyemi, T. (2025). Advancements in Artificial Intelligence in Noninvasive Cardiac Imaging: A Comprehensive Review. Clinical Cardiology, 48(1), e70087.

[10] Bednarek, A., Gumiężna, K., Baruś, P., Kochman, J., & Tomaniak, M. (2025). Artificial Intelligence in Imaging for Personalized Management of Coronary Artery Disease. Journal of Clinical Medicine, 14(2), 462.

[11] Singh, A., Nagabhooshanam, N., Kumar, R., Verma, R., Mohanasundaram, S., Manjith, R., & Rajaram, A. (2025). Deep learning based coronary artery disease detection and segmentation using

ultrasound imaging with adaptive gated SCNN models. Biomedical Signal Processing and Control, 105, 107637.

[12] Rajeev, C., & Natarajan, K. (2025). Coronary artery disease classification using ConvMixer based classifier from CT angiography images. PeerJ Computer Science, 11, e2771.

[13] Ried, I., Krinke, I., Adolf, R., Krönke, M., Moosavi, S. M., Hendrich, E., ... & Hadamitzky, M. (2025). Incremental diagnostic value of coronary computed tomography angiography derived fractional flow reserve to detect ischemia. Scientific Reports, 15(1), 12817.

[14] Pezel, T., Toupin, S., Bousson, V., Hamzi, K., Hovasse, T., Lefevre, T., ... & Garot, J. (2025). A Machine Learning Model Using Cardiac CT and MRI Data Predicts Cardiovascular Events in Obstructive Coronary Artery Disease. Radiology, 314(1), e233030.

[15] Mirjalili, F. S., Baghiani, T., Badkoubeh, F., Andishmand, A., Sarebanhassanabadi, M., Mohammadi, H., ... & Seyedhosseini, S. M. (2025). Noninvasive prediction of coronary artery disease severity: Comparative analysis of electrocardiographic findings and risk factors with SYNTAX and Gensini score. Science Progress, 108(1), 00368504241309454.