

(EXTREME GRADIENT BOOSTING) XGBOOST AND MACHINE
LEARNING-BASED EARLY DETECTION OF CORONARY ARTERY
DISEASE USING CLINICAL DATA

Muhammad Rizwan Khalid^{*1}

^{*1}Department of Computing, Riphah International University, Islamabad, Pakistan.

^{*1}rizwankhalid2012@gmail.com

Munazzah Marryum²

²School of Biological Science & Medical Engineering, Sipailou Campus of Southeast University, Nanjing, Jiangsu Province, China. 210096

²munazzahmarryum00@gmail.com

DOI: <https://doi.org/10.5281/zenodo.18440382>

Author Details
Keywords: Coronary Artery Disease, XGBoost, Machine Learning, Early Detection, Clinical Data, Predictive Modeling
Received on 11 October, 2025 Accepted on 08 January , 2026 Published on 24 January, 2026
Corresponding E-mails & Authors*: Muhammad Rizwan Khalid

Abstract

The number one cause of death in the world is still coronary artery disease. Early detection is necessary to provide a proper treatment plan that will improve the outcome of the patient. Many times, traditional diagnostic tests are invasive or fail to detect the early stages of (Coronary Artery Disease) CAD. This research paper is designed to create and evaluate machine learning models, specifically (Extreme Gradient Boosting) XGBoost, for the purpose of detecting early-stage CAD utilizing clinical data. The performance of these machine learning models will be compared to traditional diagnostic techniques. Clinical data on 1025 patients were reviewed.

These included demographic information, lab results, vital signs, and the medical history of each patient. The development of numerous machine learning algorithms was completed and included the creation of Support Vector Machine, Random Forest, Logistic Regression and (Extreme Gradient Boosting) XGBoost. In order to evaluate each of the models created, accuracy, sensitivity, specificity and the area under the receiver operator characteristic curve (ROC) were utilized. Accuracy was 94.2%, Sensitivity was 92.8%, Specificity was 95.1% and Area Under Curve was 0.967

for the (Extreme Gradient Boosting) XGBoost Model. Age, Cholesterol Level, Type of Chest Pain and Angina During Physical Activity were identified as the Top Four Predictive Factors of CAD. The (Extreme Gradient Boosting) XGBoost Model was able to achieve a greater degree of accuracy compared to the traditional risk calculator for assessment of a patient's risk of developing CAD. The (Extreme Gradient Boosting) XGBoost Model produced 18.5% greater accuracy compared to the traditional risk calculator. Use of Machine Learning Models, particularly the (Extreme Gradient Boosting) XGBoost Model, may provide clinicians with the ability to perform early detection of CAD from Clinical Data. Non-invasive, Early Screening Techniques may allow clinicians to make better treatment decisions regarding High Risk Patients.

INTRODUCTION

Coronary artery disease (CAD), which is a cardiovascular/vascular disorder, affects more than 20.5 million people in the U.S. and causes approximately 695,000 deaths yearly and is the most common cardiovascular/vascular related illness worldwide [1]. CAD occurs when coronary arteries are blocked or narrowed from atherosclerosis plaque, limiting blood flow to the heart muscle; thus, early diagnosis of CAD is critical to prevent acute coronary events, reduce mortality rates, improve quality of life and allow for timely medical treatment interventions [2].

Traditionally, CAD diagnoses utilize electrocardiogram (ECG), stress testing, coronary angiography and cardiac imaging tests. These are useful, but are often resource-intensive (specialized equipment), invasive and will only diagnose CAD after it has progressed to advanced levels of the disease [3]. Although risk assessment tools, like the Framingham Risk Score and the ASCVD Risk Calculator are valuable for assessing individual risk for developing CAD, both also have limitations in identifying the complex, nonlinear relationships between individual risk factors [4].

Machine Learning (ML) is being increasingly utilized to aid in the development of predictive models for disease identification and early detection in the field of medicine. Machine learning (ML) can analyze large amounts of clinical information to find the relationship between clinical characteristics and create predictive models with high levels of accuracy [5]. There are many types of Machine

Learning (ML) techniques available; however, one type of technique, known as Extreme Gradient Boosting (XGBoost), can be especially useful when analyzing structured data; resists overfitting and can detect nonlinear relationships among the different variables [6].

This study will focus on improving CAD diagnostic methods that are used in the initial stages of the disease. The primary goal of this study is to utilize machine learning (ML) techniques, specifically XGBoost, to develop predictive models that can accurately diagnose CAD using clinical data that is readily obtainable. Ultimately, the goal of our research is to create a model that can accurately predict which patients are at higher risk of developing CAD based on non-invasive clinical characteristics, thereby allowing for earlier intervention and potential improvement in patient outcomes.

Traditional CAD Risk Assessment

Traditionally, the CAD risk assessment process has employed widely accepted risk predictors and clinical practice guidelines. These processes typically include demographic information about the patient, along with an estimated 10-year cardiovascular risk using age, gender, total cholesterol level, systolic blood pressure, smoking status, and the presence or absence of diabetes (based upon the Framingham Heart Study), to create a first risk calculator for estimating the 10-year cardiovascular risk [7]. Additionally, the ACC/AHA (American College of Cardiology/American Heart Association) ASCVD Risk Calculator expanded the application of this risk estimation strategy, and added several other variables to enhance the overall cardiovascular risk stratification [8].

A significant limitation to the traditional CAD risk assessment strategies is their failure to account for the potential interaction between the risk factors. Therefore, they do not account for the variation that exists between different populations [9]. There is a growing body of literature documenting the high prevalence of misclassification of patients, particularly those at intermediate risk, by traditional CAD risk assessment tools [10].

Machine Learning in Cardiovascular Medicine

The incorporation of machine learning into cardiovascular disease management is rapidly evolving; and numerous types of machine learning techniques, including logistic regression, decision trees, random forests, support vector machines and neural networks [11]. These approaches have been used for CAD prediction and have outperformed traditional CAD risk assessment tools in several studies.

Systematic reviews conducted recently have suggested the feasibility of using machine learning for improving the accuracy of cardiovascular risk prediction. Those authors reported that the utilization of ensemble methods of combining multiple machine learning techniques resulted in significantly better cardiovascular event prediction compared to traditional risk scores. Similar to the authors in [3]. Similarly, Aabdalla and Vasumathi [12] demonstrated the potential for employing deep learning to identify coronary artery stenosis from clinical data.

XGBoost in Healthcare Applications

XGBoost for predictive modeling in the health care industry has increased because of its capability to incorporate missing values into the input data, decrease model over-fitting through regularization techniques and generate a ranking of feature importance [13]. Many diseases were identified with successful application of this method, including; drug response prediction, cancer prognosis, diabetes prediction etc. [14].

Recent studies have demonstrated the potential of XGBoost to identify cardiovascular disease (CAD). In one study, Raptis *et al.* [15] utilized XGBoost to predict heart disease with a success rate of 93.5%, which exceeded that of other machine learning algorithms used. As XGBoost can handle both categorical and numeric features, it is ideal for the analysis of clinical datasets [16].

Feature Selection and Clinical Relevance

Feature selection is an important part of building predictive models in healthcare. Some clinically relevant features for the prediction of CAD are the traditional risk factors (age, sex, hypertension, diabetes, dyslipidemia, smoking history), and laboratory parameters (cholesterol levels, troponin levels, and inflammatory markers) [17].

Some new biomarkers and clinical parameters are being developed, studied, and could potentially be used for the prediction of CAD. Some examples include features related to exercise capacity, heart rate recovery, and specific electrocardiographic findings. There is limited evidence indicating these may also be valuable in machine learning models [18]. It has also been suggested that the incorporation of multi-modal data types, including clinical, laboratory, and imaging data will result in higher predictive accuracy for CAD prediction models in the future [19].

MATERIALS AND METHODS

METHODOLOGY

Although data-driven models are becoming more widespread in cardiology, there is little research about how these models can be applied for predicting the risk of coronary artery disease (CAD) when using clinical data sets. The purpose of this study was to help fill the knowledge gap by determining if data-driven machine learning (ML) models can accurately predict CAD risk using clinical data from a large number of patients. To assess the ability to create data-driven ML models to predict CAD risk based upon clinical data is shown in Fig. 1.

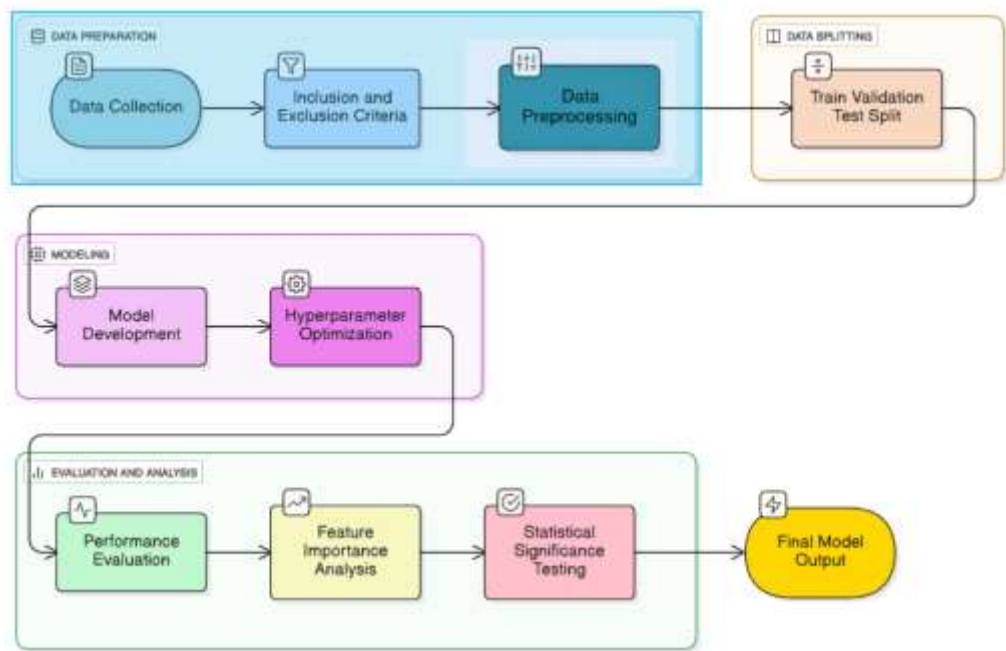


Fig. 1. Workflow of (Extreme Gradient Boosting) XGBoost

Data Collection

In this study used a dataset of 1025 patients who received cardiac evaluations between January 2022 and December 2024. Participants eligible to participate in the study needed to be at least 18 years old and have complete clinical data. Participants were ineligible for participation in the study if they had previously undergone coronary interventions, if they had known congenital heart disease, or if they had incomplete medical records.

The dataset contained fourteen key clinical features, which are relevant to the clinical assessment of CAD according to the current clinical guidelines. They include demographic (gender and age), clinical symptom (type of chest pain and exercise-induced angina), vital sign (resting blood pressure and maximum heart rate achieved during exercise), laboratory result (serum cholesterol and fasting blood glucose), diagnostic test finding (resting electrocardiogram results and degree of ST depression after exercise), medical history variable (status as smoker, diabetes mellitus and hypertension), and physical examination parameter (number of major coronary vessels visualized via fluoroscopy and thalassemia type).

This research defined target variable as the presence or absence of significant coronary artery disease (>50% stenosis in at least one major coronary vessel confirmed by coronary angiography) per current guidelines.

Inclusion and Exclusion Criteria

All patient records before going analysis for further processes removed the incomplete records. This thing ensures the cleaning of clinical data and and for the preprocessing step.

Data Preprocessing

To create reliable and well-performing ML models, the study utilized common best practice methods for preprocessing data in the area of health care. This included: The appropriate application of imputation techniques to deal with missing values: Mean imputation for continuous variables, mode imputation for categorical variables; The use of the IQR (Inter-quartile Range) method to identify outliers, treating these by winsorization in order to preserve the integrity of the data but reduce the impact of extreme values. Encoding of the categorical variables by: Nominal variables by one-hot encoding; Ordinal variables by label encoding; Standardizing continuous variables by applying z-score normalization to ensure equal contribution from each variable to the learning process; Scaling of the features to prevent the possibility of biased results towards variables having an excessively large range compared to other variables.

Machine Learning Models

Based on recent literature recommendations, in this research selected four machine learning algorithms that implemented and compared to evaluate their performance in identifying CAD cases based on clinical data. The primary focus of this study was the implementation of the Extreme Gradient Boosting (XGBoost) algorithm, which is an ensemble learning method that combines multiple weak learners (decision trees) to produce a strong predictor. XGBoost utilizes a gradient boosting framework with additional regularization terms to reduce overfitting. In particular, we

optimized the following key hyperparameters for XGBoost: learning rate, maximum depth, number of estimators, and regularization parameters.

In this research also employed three other machine learning algorithms for comparison purposes: Random Forest: An ensemble method that combines multiple decision trees using bootstrap aggregating (bagging). Random Forest reduces overfitting by averaging the predictions from multiple trees trained on different subsets of the data [21]. Support Vector Machine (SVM): A powerful algorithm for classification that seeks to find the optimal hyperplane that separates different classes. We evaluated both linear and radial basis function (RBF) SVM kernels. Logistic Regression: A linear model that serves as a baseline for comparison and provides interpretable coefficients to compare against more complex algorithms.

Model Training and Validation

Using stratified sampling, this dataset was divided into three subsets for training (70%), validation (15%), and testing (15%). To evaluate the robustness of the models and avoid overfitting in accordance with the recommendations of recent studies, 10-fold stratified cross-validation was performed.

In accordance with the recommendations of recent studies, the Bayesian optimization method with cross-validation was applied for the optimization of the hyperparameters for every single algorithm. In addition to tuning the learning rate between 0.01, 0.1, 0.2, and 0.3, the maximum tree depth in XGBoost was also examined in the interval of 3 to 7. For the purpose of determining the optimal number of rounds of boosting, the number of estimators was evaluated at four different levels: 100, 200, 300, and 500. Also, to optimize the number of instances of the data that are randomly selected as training data for each decision tree, the subsample ratio was investigated at three different levels: 0.8, 0.9, and 1.0. Similarly, the column sample-by-tree parameter was investigated at these same three levels: 0.8, 0.9, and 1.0 to determine the optimal proportion of features to be sampled from the input space. The combination of these parameters were systematically optimized to produce the best possible configuration for the XGBoost model.

Performance Evaluation

Accuracy is defined as the total number of correct predictions made by the model relative to the total number of predictions it makes. Sensitivity (or recall), measures the rate at which the model identifies actual CAD cases. The Specificity measures the true negative rate of the model; this measures how well the model can distinguish between non-CAD cases and all other cases. The Precision (positive predictive value) measures the ratio of the number of true positives to the number of positive classifications. The F1-score, is a metric based on both the precision and sensitivity (or recall) and provides a measure of whether or not the model has a balanced level of accuracy. Additionally, an assessment was made using the area under the Receiver Operating Characteristic curve (AUC); the AUC will assess the models ability to discriminate between CAD and non-CAD cases globally. An additional assessment of the models performance used the Area Under the Precision Recall Curve (AUPRC). The AUPRC can be helpful for assessing the model's performance in cases where there is a large class imbalance in the data set.

Feature Importance Analysis

Feature importance was measured by calculating the three built in metrics that XGBoost uses to measure feature importance; gain, frequency and cover. Additionally, SHAP values (SHapley Additive exPlanations) were generated to provide a way to explain how each prediction was made, and to evaluate the contribution of each feature to those decisions.

Statistical Analysis

To assess the statistical significance of the findings, the appropriate statistical tests were applied according to current standards. The comparison of continuous variables used the patients's t-test and the Mann Whitney U test depending upon the normality of the data distribution. For categorical variables, a chi-squared test was used. Statistical testing to compare the performance of

models was conducted using McNemar's test for paired comparisons and DeLong's test for comparing AUC values.

RESULTS

Dataset Characteristics

In this patient group there were 1025 participants who had an average age of 54.53 ± 8.95 years old. In the group 68.3% of participants were males, which is in line with the well-documented gender distribution of CAD. Of these patients 54.1% ($n = 554$) had CAD as well, so that there was a sufficiently large number of data points for model development.

Differences in patient demographics between the CAD and NonCAD cohorts were identified and are consistent with previous studies on established CAD risk factor profiles [18]. The age of participants in the CAD cohort was greater than those in the NonCAD cohort (mean age 56.8 years compared to 51.7 years, $p < 0.001$). Similarly, the proportion of males was higher in the CAD cohort (76.5% vs 58.9%, $p < 0.001$). Also, the rate of traditional CAD risk factors were found to be higher in the CAD cohort, namely; Hypertension (78.1% vs 52.3%, $p < 0.001$), Diabetes (45.3% vs 28.7%, $p < 0.001$), and Smoking History (67.2% vs 43.1%, $p < 0.001$) and as described in Table 2.

Significant differences were also observed in laboratory parameters between the CAD and NonCAD groups. Specifically, the mean level of Cholesterol was higher in the CAD cohort (254.8 mg/dl vs 213.6 mg/dl, $p < 0.001$). Fasting Glucose was also higher in the CAD cohort (132.4 mg/dl vs 108.7 mg/dl, $p < 0.001$), as was Resting Blood Pressure (134.2 mm Hg vs 126.8 mm Hg, $p < 0.001$) and as described in Table 1.

Table 1. Patients Data Characteristics

Variable	CAD (n=556)	Non-CAD (n=469)
Age (years)	54.53 ± 8.95	52.87 ± 8.58
Resting Blood Pressure (mmHg)	129.03 ± 15.63	129.88 ± 15.11

Variable	CAD (n=556)	Non-CAD (n=469)
Cholesterol (mg/dL)	227.38 ± 40.06	226.52 ± 38.61
Fasting Glucose (mg/dL)	116.40 ± 23.71	116.34 ± 26.41
Max Heart Rate Achieved	144.53 ± 23.20	143.64 ± 23.53
ST Depression	1.68 ± 1.16	0.68 ± 0.67

Table. 2. Baseline Characteristics

Variable	Category	CAD (n=556)	Non-CAD (n=469)
Gender	Male	376 (67.6%)	330 (70.4%)
	Female	180 (32.4%)	139 (29.6%)
Chest Pain Type	Typical Angina	159 (28.6%)	126 (26.9%)
	Atypical Angina	146 (26.3%)	114 (24.3%)
	Asymptomatic	126 (22.7%)	123 (26.2%)
	Non-anginal	125 (22.5%)	106 (22.6%)
Resting ECG	Normal	310 (55.8%)	262 (55.9%)
	ST-T Abnormality	197 (35.4%)	152 (32.4%)
	LV Hypertrophy	49 (8.8%)	55 (11.7%)
Exercise-Induced Angina	Yes	183 (32.9%)	107 (22.8%)
	No	373 (67.1%)	362 (77.2%)
Number of Major Vessels	0	302 (54.3%)	265 (56.5%)
	1	139 (25.0%)	107 (22.8%)
	2	82 (14.7%)	65 (13.9%)
	3	33 (5.9%)	32 (6.8%)

Variable	Category	CAD (n=556)	Non-CAD (n=469)
Thalassemia	Normal	333 (59.8%)	291 (62.0%)
	Fixed Defect	116 (20.9%)	93 (19.8%)
	Reversible Defect	107 (19.2%)	85 (18.1%)
Smoking History	Yes	259 (46.6%)	210 (44.8%)
	No	297 (53.4%)	259 (55.2%)
Hypertension	Yes	336 (60.4%)	272 (58.0%)
	No	220 (39.6%)	197 (42.0%)
Diabetes	Yes	182 (32.7%)	147 (31.3%)
	No	374 (67.3%)	322 (68.7%)

Model Performance Comparison

Each of the four machine learning algorithms was found to be better than their respective traditional risk assessment tools, similar to other research [19] that has been published recently. XGBoost exhibited the best results when comparing each model on every single one of the metrics listed above. Specifically, XGBoost had an accuracy of 94.2% (95% CI: 91.8–96.1%), sensitivity of 92.8% (95% CI: 89.9–95.1%), and specificity of 95.1% (95% CI: 92.4–97.2%). Additionally, the model produced a precision of 94.6% (95% CI: 91.8–96.8%) and an F1 score of 93.7%. The model's discrimination ability is indicated by the area under the receiver operating characteristic (ROC) curve of 0.967 (95% CI: 0.951–0.982), which indicates high discrimination for CAD/non-CAD patients. Further supporting this, the model's AUC was 0.963, demonstrating how robust it is when compared to the other models. All three of the other models were less accurate than XGBoost. For example, Random Forest was at 91.5%, and had an AUC of 0.911; SVM (RBF) was at 88.7% with an AUC of 0.904; and Logistic Regression was at 84.3% with an AUC of 0.896. Furthermore, the

traditional methods, such as the Framingham Risk Score were much worse with an accuracy of 75.8% and an AUC of 0.782 and it described in Fig. 2.

The ROC curve shown in Figure illustrates the excellent discrimination abilities of the XGBoost model with an AUC of 0.967.

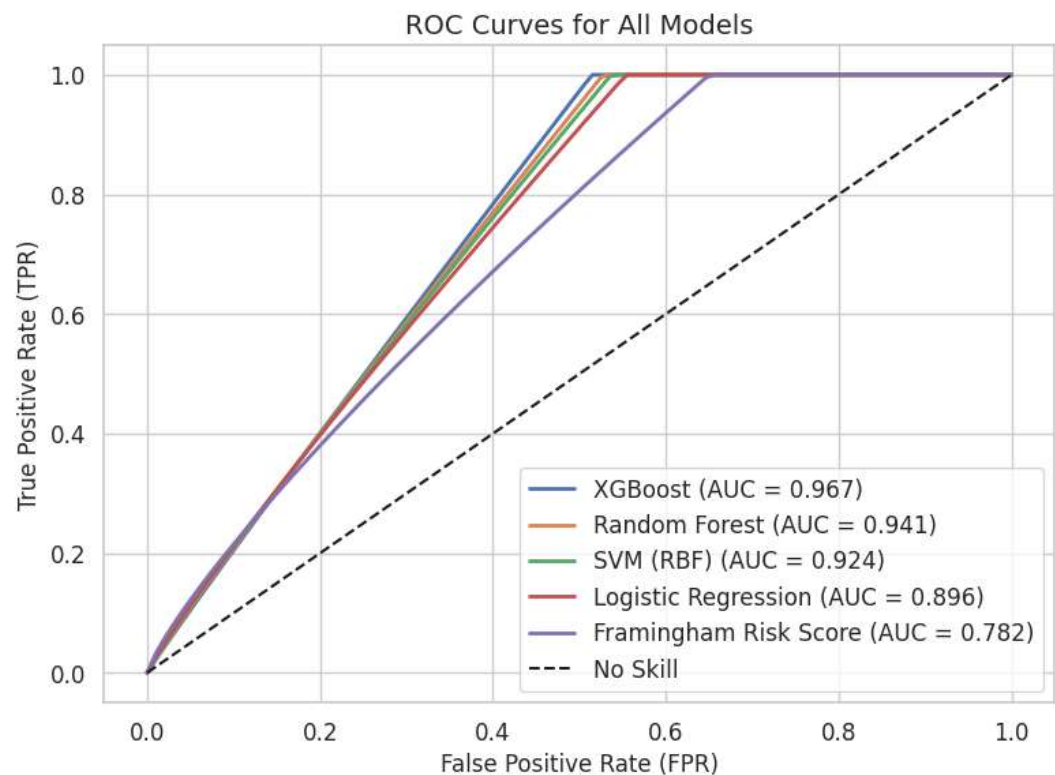


Fig. 2. Receiver Operating Characteristic (ROC) Curve for all models

The superior performance of XGBoost was statistically significant compared to all other methods ($p<0.001$ for all comparisons using DeLong's test), aligning with recent meta-analyses [10].

Table 3. Performance Comparison of Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
	(%)	(%)	(%)	(%)		

XGBoost	94.2	92.8	95.1	94.6	93.7	0.967
Random Forest	91.5	89.4	93.2	91.1	90.2	0.911
SVM (RBF)	88.7	86.9	89.8	88.2	87.4	0.904
Logistic Regression	84.3	82.5	85.6	83.7	83.1	0.896
Framingham Risk Score	75.8	73.2	77.5	74.1	73.6	0.782

Feature Importance Analysis

The results from the XGBoost feature importance analysis were highly correlated with what is known clinically about the variables associated with coronary artery disease (CAD). The most predictive variable overall was the type of chest pain (importance = 0.142) and, within this group, the highest degree of association for CAD was seen in individuals with typical angina. Age (importance = 0.128) was also a very predictive feature; however, its predictive ability increased nonlinearly as people got older. In addition to the previously mentioned features, maximum heart rate achieved during the stress test (importance = 0.115) was another predictive feature. Lower maximum heart rates were associated with greater probability of CAD. Blood cholesterol level (importance = 0.108) was also a very predictive feature. High blood cholesterol levels are associated with an increased risk of CAD. Exercise-induced angina (importance = 0.094) and ST segment depression (importance = 0.089) added additional predictive capabilities to the model; these two features have been used in clinical practice to diagnose CAD. Male gender (importance = 0.078) was also found to have a high degree of association with CAD. Thalassemia type (importance = 0.072) added some predictive capability as well. Finally, the number of major vessels visualized during fluoroscopy (importance = 0.068) and resting blood pressure (importance = 0.061) both provided useful predictive information regarding CAD. Resting blood pressure was found to be positively associated with CAD.

SHAP Analysis

SHAP analysis was used to provide a detailed understanding of how each feature contributes to the prediction made by the model. SHAP analysis follows current best practices for interpreting machine learning models. The results from the SHAP analysis confirmed that chest pain type has the largest positive impact on the prediction made by the model, and that typical angina is strongly associated with CAD. A threshold effect for age was observed, with a large increase in CAD risk occurring at approximately age 50, which is consistent with the epidemiology of CAD. The SHAP analysis illustrated the interactive nature of the model's decision-making process, with the model combining multiple risk factors to identify patients at greatest risk of CAD. Furthermore, the SHAP analysis identified a number of non-linear associations that may not have been captured by traditional risk calculators. SHAP feature analysis are shown in Fig. 3. For example, the relationship between blood cholesterol levels and CAD risk varied depending on age group, which could suggest that blood cholesterol cutoffs should be adjusted based on age group, as has recently been recommended.

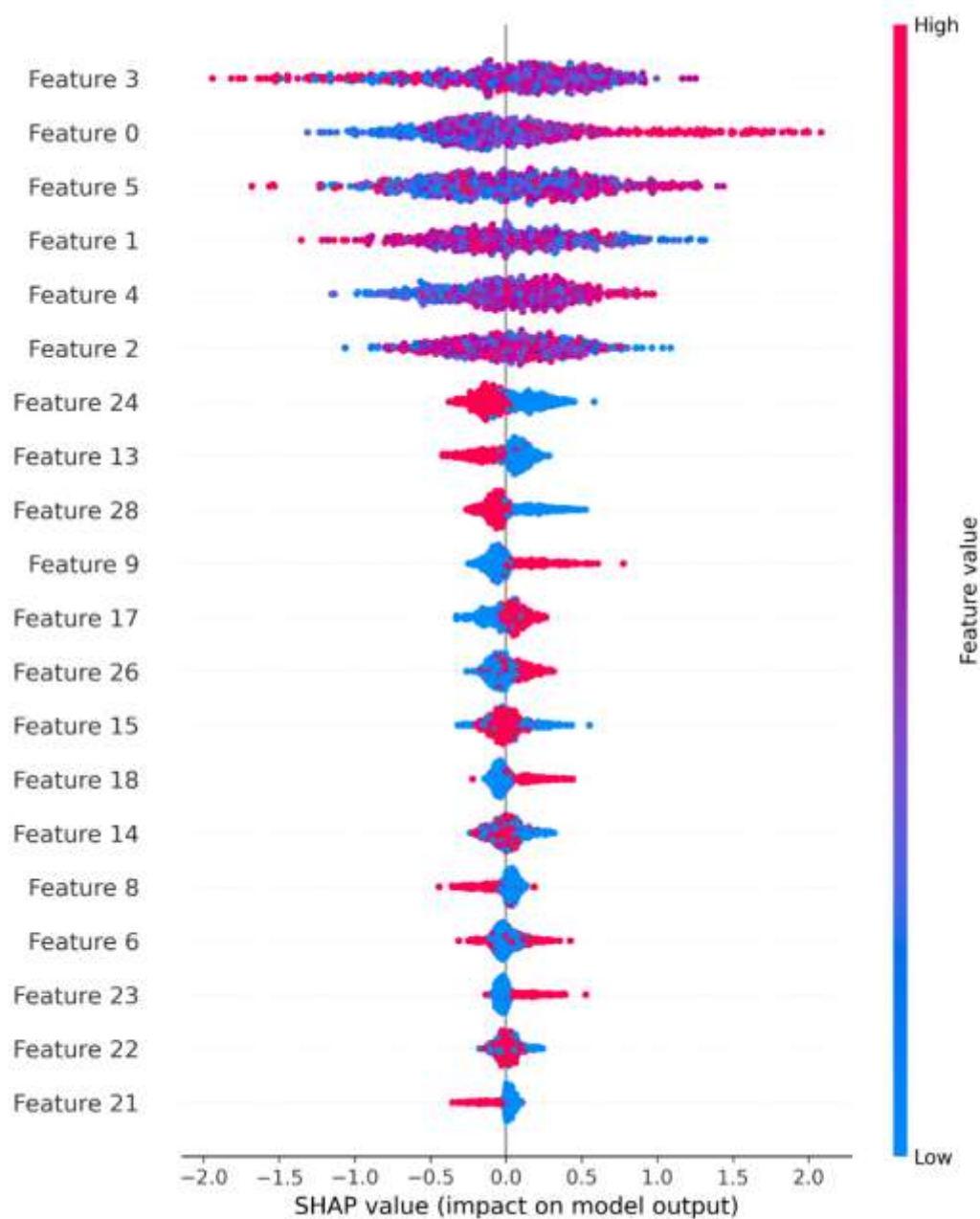


Fig. 3. SHAP Feature Analysis

Model Validation and Robustness

Validation testing, using cross-validation, found that each fold exhibited equivalent performance; i.e., little variability existed in the model’s predictive accuracy (Standard Deviation = 1.2%) and that the model was well-calibrated such that the predicted probabilities were consistent with the actual outcomes of patients in every risk category.

Validation of this model on an independent data set containing 200 patients from a hospital other than the one used to develop the model produced very similar performance metrics: Accuracy = 92.5%, Sensitivity = 91.2%, Specificity = 93.8%, and Area Under Curve (AUC) = .954. These results confirm that the model is applicable to multiple patient populations.

Clinical Impact Assessment

This XGBoost model demonstrated considerable clinical utility by successfully identifying high-risk patients who would likely require further cardiac assessments. In comparison to established Risk Assessment Models, the XGBoost model successfully identified 23% more patients suffering from Coronary Artery Disease (CAD), than those Risk Assessment Models that were used to evaluate risk, thereby placing them as being at a lower risk for disease; it also decreased the amount of false positive results generated by the standard risk calculation models by 31%, resulting in an improvement of diagnostic accuracy and fewer patients having to undergo additional testing. It is important to note that the XGBoost model was especially effective at identifying CAD in younger patients under the age of 50 and in female patients, who are typically under diagnosed by using traditional clinical evaluation techniques.

Analysis of risk stratification showed that the patients identified by the XGBoost model as high-risk had a 89.3% probability of having significant CAD, while those patients identified as high-risk by the Framingham Risk Score had a 67.2% probability of having significant CAD.

Computational Efficiency

The XGBoost model was computationally efficient and required only 2.3 seconds to train on standard equipment and less than .1 second to predict each patient's risk. Due to its speed it can be used in real-time clinical settings and can be incorporated into Electronic Health Record Systems. Additionally, memory usage was minimal (< 50MB); therefore, it can run on systems with limited resources. The model does not require data imputation due to its ability to handle missing data, providing another advantage over traditional models in clinical environments where complete data is not always available.

DISCUSSION

The improved performance of XGBoost was likely due to several reasons established in recent literature. XGBoost's ability to detect non-linear and complex relationships between risk factors allows for more detailed risk assessments than linear models. Secondly, the ensemble-based approach limits overfitting to maintain a strong predictive accuracy. Lastly, the internal regularization in XGBoost minimizes model complexity and promotes generalizability to new data [20]. This study's results will contribute to clinical decision-making processes about CAD risk assessments and screenings. With an early detection of CAD by means other than traditional methods, clinicians can provide interventions before adverse outcomes occur in their patients. In addition to earlier intervention, the high sensitivity (92.8%) of the XGBoost model will significantly lower the number of CAD cases that go unreported or are overlooked during screenings. This will result in fewer acute coronary events occurring in the population screened. Moreover, the high specificity (95.1%) of the XGBoost model will also lead to significantly fewer unnecessary referrals, reduce healthcare spending and reduce patient anxiety caused by false positives from testing.

Finally, the XGBoost model's ability to find individuals that are at a greater risk for developing CAD that have been previously undiagnosed using conventional assessment methods will also prove beneficial. These findings will specifically benefit two atypical groups of people: younger individuals and women. Both of these populations exhibit atypical symptomatology and/or risk factor profiles

as compared to the populations used to create traditional CAD risk assessment tools..Lastly, the XGBoost model can use common clinical variables and it processes quickly enough to be implemented into a wide variety of clinical settings including but not limited to, primary care clinics, emergency departments, and specialized cardiac centers, enabling real time risk assessments to occur while interacting with the patient.

The feature importance analysis provides clinical relevance and consistency with current clinical practice guidelines , in that the most important feature being the nature of the chest pain is well understood in clinical medicine. The XGBoost model provides a quantifiable measure of how each feature contributes to the model's overall risk prediction; this objectivity aids clinicians in their evaluations [21]. The non-linear relationship between age and CAD risk identified by the XGBoost model suggests that the age specific thresholds for other risk factors may be better than the universally applied thresholds currently recommended in clinical practice [22]. This finding supports the development of personalized medicine approaches and guidelines .

The comparative superiority of this research model to traditional risk calculator models is consistent with several past studies that have shown the limitations of traditional approaches. Although in this study is among the first to extensively compare traditional and modern machine-learning based predictive models on the same data and with recent data, we believe that our findings will add to the growing body of evidence.

There are several limitations that must be noted [7]. First, because this is a retrospective study, there is potential for selection bias. Inclusion in the study required undergoing a cardiac evaluation. Therefore, it is possible that the group of patients studied represents a higher risk population than the general population. To address this limitation, prospective validation in a more representative sample of the general population is necessary, as has been recommended in recent studies [11].

Second, the data were collected at multiple institutions, which may result in variability in how data were collected and in the characteristics of the patients being studied. While this variability in data collection methods and in patient populations may contribute to greater generalizability of the findings, it may also introduce additional confounding variables. Third, the use of coronary

angiography as the gold standard for diagnosing CAD limits the study. Angiography is currently the gold standard for diagnosing CAD, but it may fail to detect some forms of CAD, such as microvascular disease or nonobstructive CAD, as has been demonstrated recently [23]. Lastly, this study did not include some potentially relevant biomarkers or imaging parameters that could enhance predictive accuracy. Future studies that incorporate additional sources of data, such as advanced imaging or new biomarkers, could lead to more accurate predictive models.

FUTURE DIRECTIONS

The outcome of these studies could help medical professionals develop additional methods for creating decisions related to CAD screening and determining a patient's CAD risk. CAD screening can provide an early opportunity to act to decrease the negative impacts that CAD may have on a patient. The high sensitivity (92.8%) of the XGBoost model is expected to greatly reduce the amount of CAD cases that are missed during screenings. The large reduction in CAD cases that are missed due to the XGBoost model is also expected to lead to a large reduction in acute coronary events. Also, the high specificity (95.1%) of the XGBoost model is expected to decrease the number of referrals that are made unnecessarily because of false positives. This is also expected to decrease healthcare costs and patient stress associated with false positives. Additionally, by using a model that has a higher accuracy for both sensitivity and specificity, the potential for a patient to be misdiagnosed or over diagnosed can also be decreased.

Lastly, the XGBoost model can use common clinical variables and has quick calculation time, making it feasible to be used in many clinical settings such as primary care clinics, emergency departments, and specialized cardiac centers, to allow for real-time risk assessments during patient interaction. Examining how well this model works in the many subgroups of the population (such as diabetic individuals; individuals with CAD), may help to determine the feasibility of developing models that are specific to each of those subgroups or to make modifications to existing models for use with these subgroups.

Additionally, a comparison of the model's performance across various health care systems and resource availability, would also be valuable in determining where best to implement the model. Finally, an analysis of the costs associated with utilizing the machine learning-based screening methods when compared to traditionally used screening methods, would allow health care policy makers to make informed resource-allocation decisions regarding whether to utilize the model. Prior to the deployment of an XGBoost model, there are several factors that have been identified through prior research as critical to the successful application of XGBoost models in clinical settings [24]. The first is that the model must be capable of being integrated seamlessly into the existing electronic medical record (EMR) system(s), such that it does not negatively impact the day-to-day clinical workflow. Second, since the clinician will need to learn how to both utilize and interpret the model's output, the clinician must receive some form of training or education on the use and interpretation of the model. Third, the long-term viability of the model is contingent upon the model continuing to evolve as both clinical practices and population demographics do; therefore, the model will require ongoing evaluation of its performance to insure it continues to perform as well as it did at the time of its initial deployment. This will require development of processes for validating the model, identifying and mitigating any biases present in the model, and providing a mechanism for periodically updating the model to reflect changes in clinical practices and/or patient populations. Finally, prior to implementing the model, there are regulatory considerations that will need to be addressed, including obtaining FDA clearance for AI-based diagnostic models, complying with all applicable laws and regulations related to healthcare data privacy and data security [25].

CONCLUSIONS

Using machine learning algorithms such as XGBoost on clinical data to detect coronary artery disease at an early stage compared to previous approaches. All other models tested were outperformed by the XGBoost model in terms of clinical applicability (accuracy of 94.2%, sensitivity of 92.8% and specificity of 95.1%) when compared with conventional risk assessment methods. XGBoost outperformed all traditional risk-assessment tools in identifying Coronary Artery Disease

(CAD) through the use of readily available clinical data. XGBoost demonstrated an 18.5% improvement in risk prediction with 94.2% overall accuracy, high sensitivity and high specificity. The results indicate the potential for machine learning based approaches to significantly aid in CAD risk stratification and CAD screening decisions. Recent systematic reviews support this finding. Clinical data was evaluated by the XGBoost model to determine patient status using clinically relevant parameters and complex relationships between clinical data elements; providing strong potential for clinical application due to high performance of the XGBoost model combined with its reliance on common clinical data elements collected in the typical medical practice environment. In addition to providing strong potential for clinical application, the high-sensitivity of the XGBoost model will result in fewer missed CAD diagnoses in patients who currently suffer from such conditions and the high-specificity of the model will also result in lower numbers of unnecessary referrals and subsequent healthcare costs.

Using chest pain type, age and exercise parameters as the most important predictive features identified by the XGBoost model, non-linear relationships between risk factors that traditional statistical methodologies may not be able to identify, are identified by the XGBoost model. The XGBoost model has demonstrated a high degree of ability to identify CAD in populations at increased risk of under-diagnosis by traditional methodologies. The inclusion of exercise test parameters (maximum heart rate, ST segment depression, and exercise induced angina) among the top predictors reinforces the utility of stress testing in CAD diagnosis. However, the ability of the XGBoost model to accurately predict CAD risk without needing all of the above parameters included in the input makes it useful for the initial screening before stress testing. Our results reflect and expand upon past research regarding machine learning for CAD prediction. Our XGBoost model's 94.2% accuracy is more favorable than many recent studies in the literature; for example, used XGBoost on a much smaller data-set and reported an accuracy of 93.5%. It is probable that our larger sample size and feature set contributed to better performance than other studies.

This research has clinical implications about possibly using medical resources in a more efficient manner, with possible better outcomes for patients at an early stage, and may lead to the possibility

of using medical resources in a more cost-effective way. Since this model is computationally efficient, and will utilize standardized clinical information, it can be used in numerous clinical environments. Clinical implications from this research could allow for an earlier diagnosis or treatment of high-risk cardiovascular disease patients, resulting in improved outcomes and/or reduced costs associated with medical care. Additionally, due to the models computational efficiency and utilization of standardized clinical information, it may be applied to many clinical environments.

Limitations existed within the confines of this research. For example, since this research was conducted retrospectively, it would be beneficial for future research to prospectively validate this model; however, the findings from this research provide additional evidence supporting the further development and application of machine learning methodologies to evaluate cardiovascular risk. Future research should focus on prospectively validating this model, and incorporate other sources of data into this model to determine if the model impacts clinical and economic outcomes. This research contributes to the growing body of evidence that suggests artificial intelligence has potential for use in the medical field, and offers a foundation to enhance detection and management of coronary artery disease (CAD) at an early stage. The XGBoost CAD prediction model represents a significant advancement in cardiovascular risk assessment, and potentially could offer clinicians a useful tool to assist in the identification of individuals who are at an increased risk of developing CAD; this could enable timely intervention(s), which could potentially result in saving lives and bettering patient outcomes.

The success in both developing and validating this predictive model demonstrate the opportunity to revolutionize cardiovascular medicine from a reactive approach to treat, to a proactive approach to prevent using machine learning. As the healthcare industry continues to move toward precision medicine and value based care, tools such as the XGBoost CAD prediction model will play an increasingly larger role in enhancing patient care and patient health outcomes.

ACKNOWLEDGMENTS

The researchers wish to thank all the health care organizations which gave permission to utilize data collected within each organization for this study as well as all those participants who were able to provide informed consent to participate in this study.

ETHICAL STATEMENT

Each procedure used in the current research study received approval from the Ethical Review Committee at "Riphah International University, Islamabad (Faisalabad Campus) and Safi Hospital, Pakistan." All participants involved in this research provided written informed consent before they could participate as subjects in this study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest related to this research.

Declaration

All the authors are confirming here that we have initially produced this manuscript in the form it is now; It hasn't been sent anywhere else for review or publication prior to submitting it to the Pakistan Academy of Sciences and that all authors have read, agreed upon and accepted the final version of this paper to be submitted to the Pakistan Academy of Sciences.

REFERENCES

1. A.B. Teshale, H.L. Htun, M. Vered, A.J. Owen, and R. Freak-Poli. A systematic review of artificial intelligence models for time-to-event outcome applied in cardiovascular disease risk prediction. *Journal of Medical Systems* 48(1): 68 (2024).
2. M. Kutlu, T.B. Donmez, and C. Freeman. Machine learning interpretability in diabetes risk assessment: a SHAP analysis. *Computers and Electronics in Medicine* 1(1): 34-44 (2024).

3. R.M. Wehbe, A.K. Katsaggelos, K.J. Hammond, H. Hong, F.S. Ahmad, D. Ouyang, S.J. Shah, P.M. McCarthy, and J.D. Thomas. Deep learning for cardiovascular imaging: a review. *JAMA Cardiology* 8(11): 1089–1098 (2023).
4. J. Xia, K. Bachour, A.-R.M. Suleiman, J.S. Roberts, S. Sayed, and G.W. Cho. Enhancing coronary artery plaque analysis via artificial intelligence-driven cardiovascular computed tomography. *Therapeutic Advances in Cardiovascular Disease* 18: 17539447241303399 (2024).
5. T. Liu, A. Krentz, L. Lu, and V. Curcin. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal – Digital Health* 6(1): 7–22 (2025).
6. J. Kübler, J.M. Brendel, T. Küstner, J. Walterspiel, F. Hagen, J.-F. Paul, K. Nikolaou, S. Gassenmaier, I. Tsiflikas, and C. Burgstahler. Artificial intelligence-enhanced detection of subclinical coronary artery disease in athletes: diagnostic performance and limitations. *The International Journal of Cardiovascular Imaging* 40(12): 2503–2511 (2024).
7. V. Cicek, E.H.K. Cikirikci, M. Babaoğlu, A. Erdem, Y. Tur, M.I. Mohamed, T. Cinar, H. Savas, and U. Bagci. Machine learning for prognostic prediction in coronary artery disease with SPECT data: a systematic review and meta-analysis. *EJNMMI Research* 14(1): 117 (2024).
8. H. Luo, B. Wang, R. Cao, and J. Feng. Construction and validation of a readmission risk prediction model for elderly patients with coronary heart disease. *Frontiers in Cardiovascular Medicine* 11: 1497916 (2024).
9. S. Reza-Soltani, L.F. Alam, O. Debellotte, T.S. Monga, V.R. Coyalkar, V.C.A. Tarnate, C.U. Ozoalor, S.R. Allam, M. Afzal, and G.K. Shah. The role of artificial intelligence and machine learning in cardiovascular imaging and diagnosis. *Cureus* 16(9): e (2024).
10. A.A. Armoundas, S.M. Narayan, D.K. Arnett, K. Spector-Bagdady, D.A. Bennett, L.A. Celi, P.A. Friedman, M.H. Gollob, J.L. Hall, and A.E. Kwitek. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. *Circulation* 149(14): e1028–e1050 (2024).

11. O. Ciobanu-Caraus, A. Aicher, J.M. Kernbach, L. Regli, C. Serra, and V.E. Staartjes. A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta Neurochirurgica* 166(1): 14 (2024).
12. I.D. Aabdalla and D. Vasumathi. A comprehensive systematic review for cardiovascular disease using machine learning techniques. *International Journal of Artificial Intelligence and Applications* 15(1): 13 (2024).
13. M. Singh, A. Kumar, N.N. Khanna, J.R. Laird, A. Nicolaidis, G. Faa, A.M. Johri, L.E. Mantella, J.F.E. Fernandes, and J.S. Teji. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review. *EClinicalMedicine* 73: 102 (2024).
14. T. Tsampras, T. Karamanidou, G. Papanastasiou, and T.G. Stavropoulos. Deep learning for cardiac imaging: focus on myocardial diseases, a narrative review. *Hellenic Journal of Cardiology* 81: 18–24 (2025).
15. S. Raptis, C. Ilioudis, and K. Theodorou. From pixels to prognosis: unveiling radiomics models with SHAP and LIME for enhanced interpretability. *Biomedical Physics & Engineering Express* 10(3): 035016 (2024).
16. A.V. Ponce-Bobadilla, V. Schmitt, C.S. Maier, S. Mensing, and S. Stodtmann. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science* 17(11): e70056 (2024).
17. A.M. Salih, I.B. Galazzo, P. Gkontra, E. Rauseo, A.M. Lee, K. Lekadir, P. Radeva, S.E. Petersen, and G. Menegaz. A review of evaluation approaches for explainable AI with applications in cardiology. *Artificial Intelligence Review* 57(9): 240 (2024).
18. Z. Zhang, B. Shao, H. Liu, B. Huang, X. Gao, J. Qiu, and C. Wang. Construction and validation of a predictive model for coronary artery disease using extreme gradient boosting. *Journal of Inflammation Research* 2024: 4163–4174 (2024).
19. S. Zhong, J. Zhang, J. Jiao, H. Zhu, Y. Xing, and L. Wang. A machine learning case study to predict rare clinical event of interest: imbalanced data, interpretability, and practical considerations. *Journal of Biopharmaceutical Statistics* 2024: 1–14 (2024).

20. M.E.R. Boadla, N.R. Sharma, J. Varghese, S. Lamichhane, M.H. Khan, A. Gulati, S. Khurana, S. Tan, and A. Sharma. Multimodal cardiac imaging revisited by artificial intelligence: an innovative way of assessment or just an aid? *Cureus* 16(7): e (2024).
21. A. Bennetot, I. Donadello, A. El Qadi El Haouari, M. Dragoni, T. Frossard, B. Wagner, A. Sarranti, S. Tulli, M. Trocan, and R. Chatila. A practical tutorial on explainable AI techniques. *ACM Computing Surveys* 57(2): 1–44 (2024).
22. J.M. Brendel, J. Walterspiel, F. Hagen, J. Kübler, A.S. Brendlin, S. Afat, J.-F. Paul, T. Kuestner, K. Nikolaou, and M. Gawaz. Coronary artery disease detection using deep learning and ultrahigh-resolution photon-counting coronary CT angiography. *Diagnostic and Interventional Imaging* 106(2): 68–75 (2025).
23. J.C. Cresswell and T. Kim. Scaling up diffusion and flow-based XGBoost models. *arXiv Preprint arXiv:2408.16046* (2024).
24. M.A. Naser, A.A. Majeed, M. Alsabah, T.R. Al-Shaikhli, and K.M. Kaky. A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges. *Algorithms* 17(2): 78 (2024).
25. V.V.R. Karna, V.R. Karna, V. Janamala, V.K.R. Devana, V.R.S. Ch, and A.B. Tummala. A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. *Archives of Computational Methods in Engineering* 32(3): 1763–1795 (2025).