Crimson Publishers
Wings to the Research

Research Article

Comparative Analysis of Machine Learning Algorithms for Predicting the Risk of Recurrent Coronary Artery Disease within a 6-Month Post-Treatment Window

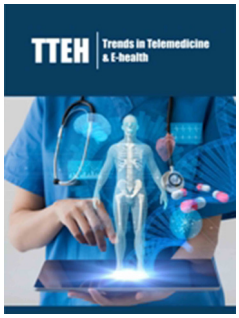
Karamo Bah[1], Adama Ns Bah[1], Wurry Jallow A[2] and Musa Touray[3]*

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan

[2]Department of Medical Laboratory Science and Biotechnology, Taipei Medical University, Taiwan

[3]School of Medicine and Allied Health Sciences, University of the Gambia, West Africa

ISSN: 2689-2707



***Corresponding author:** Musa Touray, School of Medicine and Allied Health Sciences, University of the Gambia, MDI Road, Kanifing P.O. Box 3530, Serrekunda, West Africa

Submission: February 15, 2024
Published: February 29, 2024

Volume 4 - Issue 5

How to cite this article: Karamo Bah, Adama Ns Bah, Wurry Jallow A and Musa Touray*. Comparative Analysis of Machine Learning Algorithms for Predicting the Risk of Recurrent Coronary Artery Disease within a 6-Month Post-Treatment Window. Trends Telemed E-Health. 4(5). TTEH. 000600. 2024.
DOI: 10.31031/TTEH.2024.04.000600

Abstract

Background: Cardiovascular diseases, particularly Coronary Artery Disease (CAD), remain the leading cause of death worldwide, imposing significant health and economic burdens. It is crucial to emphasize early diagnosis of CAD to prevent complications and improve patient outcomes. This study aims to predict the likelihood of CAD recurrence within 6 months post-treatment.

Methods: The Medical Information Mart for Intensive Care (MIMIC-III) database was used to perform a retrospective study. Predictive features include demographic data and laboratory test results. A 6-month CAD recurrence was set as the study outcome. We used the Machine Learning (ML) Methods Of Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) to develop a predictive model for CAD recurrence. The prognostic capacity and clinical utility of these three models were compared using the Area Under the Receiver Operating Characteristic Curves (AUROC), precision, sensitivity, specificity, f1 measure and Area Under Precision -Recall (AUPR) curve.

Results: Of 7,583 CAD patients in this study population, 2,361 (31%) had CAD recurrence during 6-month follow-up. Out of 38 features selected and extracted from the MIMIC III database, 15 variables were chosen using stepwise regression. The RF model performed best with an AUC of 0.83. The top 6 significant features in our model were platelet, WBC, RBC, INR, chloride, and creatinine.

Conclusion: Our study shows that the random forest model outperforms the XGBoost and LR models in predicting CAD recurrence within 6 months post-treatment. The study suggests a connection between certain lab indices (platelet count, WBC, RBC, INR, chloride, calcium, creatinine) and CAD recurrence, bridging knowledge gaps and guiding future research on preventive strategies and treatments for CAD.

Keywords: Prediction; Machine learning; Coronary artery disease; Recurrence; MIMIC-III database

Introduction

Coronary Artery Disease (CAD) is recognized as the leading cause of death worldwide, affecting approximately 1.72% of the global population and resulting in 9 million mortality cases per year. In the United States (US) and Europe, CAD prevalence for adults was estimated at 7.1% from 2017 to 2020 [1] and 5.11% in 2019, respectively [2]. According to the Global Burden Of Disease (GBD) study, the global prevalence of CAD was 154 million in 2016, accounting for 32.7% of the global burden of cardiovascular disease and 2.2% of the overall global burden of disease (GBD Disease Injury Incidence Prevalence Collaborators). Based on data from a national health survey conducted from 2009 to 2012, the American Heart Association (AHA) estimated a CAD prevalence of about 15.5 million, with 7.6% of men and

5.0% of women in the USA living with CAD during that period. The ONACI registry in France reported CAD incidence rates ranging from ~1% per year among men aged 45-65 to ~4% in patients aged 75-84, regardless of sex [3].

The CAD's etiological risk factors can be broadly classified into non-modifiable and modifiable factors. Non-modifiable risk factors include gender, genetics, age and family history and modifiable risk factors include smoking, obesity, lipid levels, homocystinuria and psychosocial stress, hypertension, cigarette smoking, diabetes, and physical inactivity. Recently, a faster-paced lifestyle has led people to eat more fast foods and unhealthy meals, leading to an increased prevalence of CAD [4]. The common pathological process that causes CAD is atherosclerosis, an inflammatory disease of the arteries associated with lipid deposition and metabolic alterations due to multiple risk factors. More than 70% of at-risk individuals have multiple risk factors for CAD, and only 2%-7% of the general population have no risk factors [5]. Despite the search for novel risk factors for CAD, established modified risk factors still play a major role [6]. These are associated with an increased risk in major prospective epidemiological studies [7].

To measure the world impact of CAD requires estimating CAD mortality, prevalence, and disability for men and women, by age and different regions in the world. Most of the time, nonfatal CAD incidence and prevalence are not always associated with CAD mortality. For example, improved acute and chronic CAD treatments might minimize CAD mortality and a growing population of chronic CAD survivors. Conversely, even if CAD incidence is high, high case fatality may lead to relatively low prevalence. Regardless of the time trend in age-standardized CAD prevalence, population growth and old age may increase the absolute number of people living with nonfatal CAD [8]. An increasing number of individuals with non-fatal CAD live with chronic disabilities and impaired quality of life [9]. The increasing incidence of CAD is expected to continue, due not only to the increased prevalence of obesity, diabetes, and metabolic syndrome but also to population ageing [10]. The past two decades have witnessed a steep rise in global population ageing [11]. Indeed, the United Nations estimates an increase in the population aged over 65 years from one in 11 in 2019 to one in six by 2050 [12]. Emerging issues with social relationships, psychological distress, and less than six hours of sleep a night also contribute to CAD in the current generation [12]. Data from the National Health and Nutrition Examination Survey (NHANES) from the period between 2003 to 2006 stated that an estimated 17.6 million Americans aged 20 or older had CAD, with an overall prevalence of 7.9% [10].

Machine Learning (ML) is currently one of the hot topics. It is a field of computer science that uses computer algorithms to identify patterns in huge datasets with multiple variables and can predict various outcomes based on given data. Machine learning algorithms typically split the data into training and testing sets. The model is built using the training data, and predictions and data-driven decisions are made using the testing data. ML methods have recently emerged as highly effective tools in various disciplines, including internet search engines, natural language processing, finance, healthcare, business, economics, and robotics [13]. The significance of utilizing ML models with data from MIMIC to predict prognostic outcomes has been studied on various scales. Therefore, this study aims to evaluate the risk indicators and develop a predictive model that can estimate the likelihood of CAD recurrence during a 6-month follow-up period after treatment. The findings may contribute to improved treatment planning and preventive measures for patients with CAD.

Literature Review

Coronary Artery Disease (CAD) arises when the myocardium receives an insufficient supply of oxygen and blood, stemming from the occlusion of coronary arteries. This discrepancy between oxygen demand and supply takes shape due to the formation of plaques within the luminal space of these arteries, thus obstructing the natural blood flow. This condition, once uncommon as a cause of death, has evolved into a significant global health concern. Throughout the 20th century, it gradually emerged as a primary contributor to mortality, peaking in the mid-1960s before a subsequent decline. Nevertheless, despite advancements in medical understanding and treatment, CAD remains a pervasive and leading cause of death worldwide [14]. In recent years, the integration of Machine Learning (ML) methods has significantly revolutionized the realm of disease detection and diagnosis [15,16]. Broadly, ML strategies involve the 'training' of algorithms using a reference dataset where the disease status (presence or absence) is established. Subsequently, these trained algorithms are deployed on diverse datasets to predict the disease status for patients whose condition is undetermined. As datasets expand in size, ML algorithms progressively refine their predictive capabilities, bolstering their role as disease predictors. Leveraging ML for enhanced disease prediction empowers clinicians with superior tools for detection, diagnosis, classification, risk stratification, and patient management, potentially reducing the need for extensive clinical intervention.

According to Pooja et al. [17], machine learning has demonstrated potential in identifying cardiovascular ailments based on patients' clinical data. Among the models, the random forest model exhibited remarkable accuracy, registering at 86.60%. Forssen & colleagues [18] conducted a study to assess the predictive capabilities of three distinct machine learning algorithms; logistic regression, Principal Components Analysis (PCA), and random forest, regarding the occurrence of CAD. Utilizing data from the Clinical Cohorts in Coronary Disease Collaboration (4C), which was compiled from UK NHS hospitals, their investigation focused on cases where CAD was determined based on the presence of over 50% stenosis in multiple coronary arteries. The results indicate that the random forest model exhibited superior performance with both a higher AUC (0.675) and accuracy (0.713) compared to the logistic regression model with PCA-derived features, which achieved an AUC of 0.625 and an accuracy of 0.686. However, in the case of adjusted models, the adjusted logistic regression model outperformed the adjusted

random forest model, boasting an AUC of 0.767 and an accuracy of 0.759, while the adjusted random forest achieved an AUC of 0.711 and an accuracy of 0.732.

In a separate study by Motarwar & colleagues [19], an algorithmic framework was devised to predict CAD risk, where the random forest model again emerged as the most precise with a notable accuracy of 95.08%. Another study [20] investigated the prediction of CAD through the utilization of metabolites and combined this with the efficacy of traditional risk factors such as lipid levels, blood pressure, lifestyle parameters, family history, sex, and age. Employing logistic regression with LASSO, they endeavored to identify relevant metabolites linked to CAD and subsequently compared the predictive prowess of these metabolites with those not correlated to risk factors. Surprisingly, their findings revealed that leveraging metabolites independent of conventional risk factors did not yield improvements in risk prediction based on traditional factors, specifically within cohorts of individuals devoid of CAD. Sabarish & Parvati [21] employed multiple algorithms, including Decision Tree (DT), Naive Bayes (NB), k-Nearest Neighbors (KNN), and RF, revealing that the KNN algorithm attained the highest accuracy rate at 90.7%. Furthermore, leveraging a CNN model, the detection of images related to Chinese herbal medication achieved a commendable 71% overall accuracy. Noteworthy, an impressive success rate of 96.7% was observed when employing ANN for the detection of lung cancer [22]. Notably, a research conducted by Pooja et al. [17] suggested the utilization of Natural Language Processing (NLP) to train and evaluate a depression prediction model [23], on the other hand, introduced a neural network for diabetes prediction, achieving an impressive prediction accuracy of 87.3%. In a comparative context, this innovative approach showcased substantial enhancements to the neural network's performance. In training, the performance was elevated by 91%, while testing witnessed a commendable 86% boost compared to baseline performances of 89% and 81%, respectively.

Materials and Methods

Data source and study population

Data from the Medical Information Mart for Intensive Care (MIMIC-III) v1.4 database for ages 15 and above were collected and included in the study. MIMIC-III is an openly available database containing de-identified data on 46,520 patients and 58,976 admissions of the Beth Israel Deaconess Medical Center, Boston, USA, between 1 June 2001 and 31 October 2012. These data include comprehensive descriptions, such as demographics, admission notes, International Classification of Diseases-9th revision (ICD-9) diagnoses, laboratory tests, medications, procedures, fluid balance, discharge summaries, vital sign measurements undertaken at the bedside, caregiver's notes, radiology reports and survival data [24]. The proportion of missing values in all the selected features was less than 10% therefore we removed all the missing values. A total of 7,583 patients diagnosed of CAD were enrolled in this study, 5,222 patients had no recurrence of CAD and 2361 patients had recurrence of CAD. The subjects were randomly sampled into a training and a testing set at a ratio of 80: 20. The training set was used to develop the model and the testing set was used to test the performance of the model after learning. The models were developed using the training set by using Python version 3.1 programming language (http://www.python.org) and the data preprocessing was done in R statistical software version 4.3.0.

Class imbalance

Out of total CAD patients enrolled in the study, 2,631 patients experienced a recurrence within a 6 month follow-up window. If we were to allocate patients randomly into either the training or testing set without considering this inherent imbalance, our algorithm could inadvertently learn to predominantly predict CAD non-recurrence (the larger class), undermining the central objective of our study. To effectively address this challenge, our study employs the Synthetic Minority Over-sampling Technique (SMOTE) procedure, as introduced by Chawla et al. [25]. This sophisticated method deviates from the traditional approach of merely duplicating existing instances. Instead, it ingeniously generates synthetic examples that capture the underlying patterns of the minority class. SMOTE helps to rectify the class imbalance issue, enabling our algorithm to more accurately learn and predict instances of CAD recurrence. SMOTE's ability to strategically expand the minority class enhances the robustness of our predictive model. This not only enables the algorithm to better capture the nuances of CAD recurrence but also aligns with the overarching research goal of our study.

Study outcome (cases and controls)

In the context of the study "Evaluating the likelihood of recurrence of CAD within 6 months post-treatment," the cases and controls are defined as follows: The cases (recurring CAD) refer to individuals who experienced a recurrence of CAD within 6 months after treatment. These patients had previously undergone CAD treatment but subsequently developed the disease again during the 6-month follow-up period. The cases are the group of interest for assessing the likelihood of CAD recurrence. In contrast, the controls (non-recurring CAD) refer to individuals who underwent the same CAD treatment but did not experience a recurrence within 6 months post-treatment. These individuals achieved disease stability or remission after the initial treatment. The controls were selected to represent patients with a similar baseline CAD diagnosis and treatment but without recurrent disease within 6 months. Including controls in the study aims to establish a comparison group that can help evaluate the association between the features and CAD recurrence. By comparing the cases and controls within the same population, we can identify any differences or patterns that may indicate the likelihood of CAD recurrence. For this purpose, cases and controls were analyzed by randomly sampling patients visit, as shown in Figure 1.

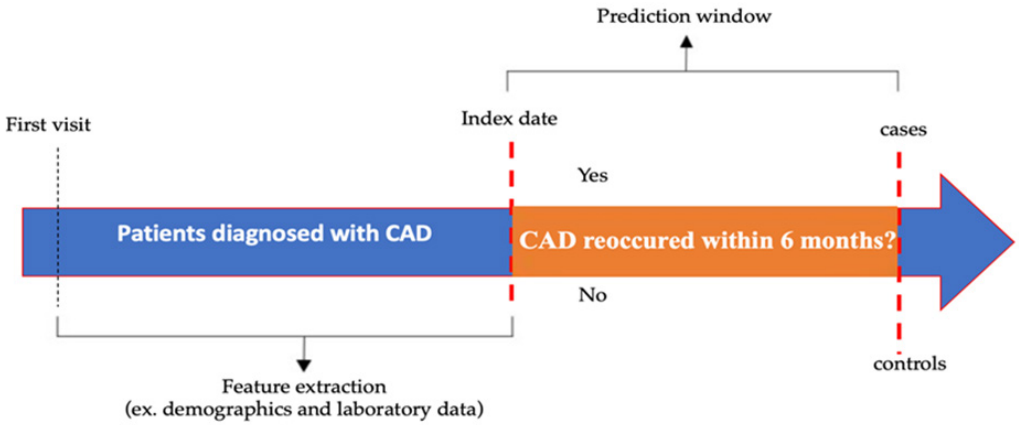


Figure 1: Timeline of study period schema.

Feature selection

The selection of predictors was a collaborative process, drawing on insights from a comprehensive literature review and a consensus meeting with a Cardiovascular Disease (CVD) specialist physician. From this groundwork, relevant features were extracted from both demographic and laboratory data sources. To discern the most influential variables among the extracted features, a stepwise logistic regression model was employed. Stepwise regression represents a dynamic statistical technique that automates the predictor selection process. Within this approach, distinct strategies are harnessed, namely forward selection, backward elimination, or a combination of the two [26]. In our study, we utilized both forward selection and backward elimination, employing the Akaike Information Criterion (AIC) as a pivotal feature selection metric. By harnessing this stepwise methodology, our study aimed to discern a subset of predictors that significantly contribute to the predictive model's performance. This approach not only leverages the physician's expertise but also employs a systematic and statistically grounded technique to refine our model's feature set. Ultimately, the selected features, informed by both clinical insights and rigorous statistical criteria, collectively enhance the model's ability to accurately predict cardiovascular disease recurrence.

Machine learning models

Within the scope of this investigation, three distinct machine learning models have been harnessed to forecast the recurrence of CAD within a critical 6-month window following treatment. These models encompass Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), and Random Forest (RF).

Logistic regression (LR): Logistic Regression (LR) is a foundational approach in predictive analytics that examines the relationship between predictor variables and the likelihood of a specific event occurring. Recently, LR analysis has become an increasingly used statistical tool in healthcare research, especially over the last two decades [27], although its origin can be dated way back to the earlier nineteenth century. It is usually considered the statistical analysis of preference when a binary (dichotomous) outcome is to be predicted from one or more independent variables. The LR is used in scenarios where we want to predict a binary outcome class (categorical) example if a specific group have a disease condition or not or where the decision is true or false.

Extreme Gradient Boosting (XGBoost) model

Extreme Gradient Boosting (XGBoost) represents an advanced ensemble learning technique known for its robustness and predictive accuracy. XGBoost constructs a complex predictive model by iteratively combining the predictive capabilities of multiple weak learners. XGBoost is an algorithm based on the gradient boosting decision tree, which can efficiently construct boosted trees and run in parallel. The boosted trees in XGBoost are divided into regression trees and classification trees. The core of the algorithm is to optimize the value of the objective function [28]. XGBoost has the advantage of scalability in all scenarios and is fast [28]. The model works by combining a set of weaker machine-learning algorithms to obtain an improved machine-learning algorithm [29]. It can identify non-linear patterns in the data. It can handle both numerical and categorical data [30], is relatively robust to outliers, provides useful estimates of variable importance and has an efficient method for estimating missing data [31].

Random Forest (RF): Random Forest is a common machine learning algorithm used for several types of classification, regression and other problems and it can model complex interactions among exploratory variables [25]. It is an ensemble of tree-structured classifiers [31] and leverages the collective use of multiple decision trees. By aggregating the predictions of individual trees, RF enhances the model's stability and predictive power. Every tree in the forest gives a unit vote, assigning each input to the most probable class label. The hyperparameters with the training set were estimated using grid search and tenfold cross-validation to find the best-optimized hyperparameters. The best parameter combination in the model was selected which led to the high performance of the models.

Model optimization

Model optimization is a crucial phase in machine learning where various aspects of a model are fine-tuned to enhance its performance and generalizability. This involves adjusting hyperparameters, selecting relevant features, preprocessing data, employing cross-validation, leveraging ensemble methods, applying

regularization, selecting appropriate algorithms, and using relevant evaluation metrics. The goal is to create a well-performing model that strikes a balance between capturing patterns and avoiding overfitting, resulting in accurate predictions on both training and unseen data. To ensure robust generalization and mitigate the potential for overfitting, we employed a hybrid approach that combines both holdout and cross-validation methodologies. Our dataset was partitioned into distinct subsets: 80% for training and 20% for independent testing. Within the training subset, we conducted a comprehensive 10-fold cross-validation procedure. The performance evaluation of our models was then conducted using the separate test subset. Notably, a grid search technique was employed during the parameter tuning process to optimize model performance.

Evaluation metrics

The models' performances were assessed using several evaluation metrics, such as the Area Under the Curve (AUC), accuracy, recall, precision, specificity, F1 measure, and precision-recall curve. These evaluation metrics collectively provide insights into different aspects of a model's performance.

A. The metrics were calculated based on the following parameters:

B. True Positive (TP): Patients who are correctly labelled by the classifier as CAD recurrence.

C. True Negative (TN): Patients who are correctly labelled by the classifier as non-CAD recurrence.

D. False Negative (FN): Patients that are incorrectly labelled by the classifier as CAD recurrence and

E. False Positive (FP): Patients that are incorrectly labelled by the classifier as non-CAD recurrence.

Below, a detailed explanation of the metrics and formula.

Receiver Operating Characteristic (ROC): The ROC analysis serves as a pivotal tool in assessing the discriminative prowess of a classification model. This technique systematically evaluates the model's ability to distinguish between different classes or outcomes. The ROC curve itself represents a graphical representation of the model's performance, with values ranging from 0.5 to 1.0. Notably, a higher ROC score corresponds to superior model performance in distinguishing between classes. This numeric representation quantifies the model's efficiency in correctly classifying instances while navigating the inherent trade-off between sensitivity and specificity.

Precision-Recall (PR) Curve: The Precision-Recall (PR) curve is constructed based on precision and recall values. Typically, recall is plotted along the horizontal axis, while precision is depicted along the vertical axis. The PR curve, alternatively referred to as average precision, quantifies performance by calculating the area under the PR curve. Especially pertinent when dealing with imbalanced datasets, the PR curve provides a valuable assessment of model effectiveness. Its ability to capture nuanced class distribution dynamics makes it particularly useful for gauging performance in situations where one class vastly outweighs the other.

Accuracy: Accuracy is a pivotal metric in the realm of classification analysis, quantifying the model's ability to correctly predict instances across all classes. It represents the ratio of correctly predicted instances to the total number of instances in the dataset [32-34].

$$\text{Accuracy} = (TP + TN)/(TP+FP+TN+FN)$$

F1 score: The F1 score, often recognized as the F1 measure or F1 statistic, emerges as a pivotal metric in the evaluation of classification models. Derived from the harmonic mean of precision and recall, the F1 score strikes a balance between these two vital performance indicators. This balance is particularly valuable in situations where class distribution is imbalanced, ensuring that both false positives and false negatives are considered. Essentially, the F1 score encapsulates the model's ability to harmonize accurate positive predictions (precision) with capturing all relevant positive instances (recall). This attribute makes the F1 score an essential tool for assessing a model's efficacy across various contexts. A high F1 score indicates that the model is achieving an optimal equilibrium between precision and recall, signifying a robust ability to both identify relevant instances and minimize erroneous classifications. Its utility extends to scenarios where striking a balance between false positives and false negatives is of paramount importance, contributing to informed decision-making and effective model selection [33].

$$\text{F-score} = (2*\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall})$$

Sensitivity (recall or true positive rate): Sensitivity, often referred to as the true positive rate or recall, is a fundamental metric in the realm of classification analysis. It measures the model's capability to correctly identify positive instances from the total instances that are actually positive. In essence, sensitivity quantifies how effectively the model "senses" or "recalls" the presence of the target class. A high sensitivity score indicates that the model is adept at identifying a significant proportion of the actual positive instances, minimizing the risk of false negatives. This is particularly crucial in scenarios where missing positive instances can have serious implications, such as in medical diagnoses [32-34].

$$\text{Recall} = TP/(TP+FN)$$

Precision (Positive Predictive Value): Precision, often termed positive predictive value, is a pivotal metric in the realm of classification analysis. It gauges the model's precision in correctly identifying positive instances from the total instances that it predicts as positive. Essentially, precision quantifies how well the model "filters" out instances that are truly positive from those that might be falsely classified. A high precision score signifies that the model is skillful at correctly labeling instances as positive, minimizing the risk of false positives. This is particularly significant in scenarios where false positives can lead to undesired consequences or unnecessary interventions. Precision serves as a critical gauge of a model's performance, especially when accuracy in labeling positive instances is crucial. In contexts where ensuring the authenticity of positive predictions is paramount, a high precision value reflects that the model is making predictions with a high level of certainty and accuracy, contributing to sound decision-

making and trustworthy outcomes.

$$Precision=TP/(TP+FP)$$

Specificity (true negative rate): Specificity, often termed the true negative rate, is a fundamental metric in the realm of classification analysis. It evaluates the model's ability to accurately identify negative instances from the total instances that are actually negative. In essence, specificity quantifies the model's "specific" ability to distinguish the absence of the target class. A high specificity score indicates that the model excels in correctly classifying negative instances, reducing the likelihood of false positives. This is particularly vital in scenarios where avoiding false alarms is crucial, such as in safety-critical applications [32-34].

$$Specificity=TN/(TN+FP)$$

Statistical analysis

Patient demographic and clinical attributes concerning recurrence and non-recurrence of CAD underwent comparative analysis. The student t-test was employed for normally distributed continuous variables, while the Wilcoxon rank-sum test was applied for non-normally distributed continuous variables. In instances of categorical variables, the Chi-square test or Fisher's exact test was utilized for inter-group differentiation. To succinctly present the data, normally distributed variables were expressed as mean and Standard Deviation (SD), non-normally distributed variables as median and interquartile range, and categorical variables as absolute values and percentages. Notably, all statistical analyses adhered to the established threshold of significance, with a p-value less than 0.05 deemed statistically significant. The entirety of data preprocessing and subsequent statistical evaluations was executed using R software version 4.3.0 (as of April 21, 2023), while the development of machine learning models was conducted employing Python version 3.1. This comprehensive approach facilitated robust analysis and model generation within a meticulously curated framework.

Results

Subject's characteristics

A total of 7,583 patients with CAD in the MIIMIC III 1.4 dataset were enrolled in this study. Out of the total study population, 2,361 (31%) had CAD recurrence. The median age for recurring CAD patients was 69 (range 31-88) years and majority were male 70% (n=1632). Patients with recurring CAD exhibit significant alterations in various blood parameters. Some of these changes include a decreased amount of white blood cells (median 9.87, IQR 7.8 - 12.2), hematocrit (median 25.4, IQR 0-32), chloride (median 103.53, IQR 100.8 - 105.87), and low-density lipoprotein (mean 8.36, SD± 0.36). Conversely, they show an increased amount of total calcium (median 8.47, IQR 8.09-8.84) and creatinine (median 1.1, IQR 0.85-1.64) (Table 1).

Table 1: Baseline characteristics of patients with non-recurrence and recurrence of CAD.

WBC: White Blood Cell; RBC: Red Blood Cell; INR: International Normalized Ratio; CK: Creatinine Kinase; HbA1c: hemoglobin A1C; LDL: Low-Density Lipoprotein; mg/dl, milligrams per deciliter; IU/L; International units per litre; cm: Centimetre; yrs: Years; S.D: Standard Deviation; K/uL: Thousand per microliter; m/uL: Million per microliter; %: percentage; mEq/L: Milli equivalents per litre. Continuous variables that are normally distributed are recorded as mean (S.D), non-normally distributed continuous variables as median (interquartile range) and categorical variables as absolute numbers and percentages, n(%). The Chi-square test was used for the comparison of categorical variables and the two-sample t-test for continuous variables. All p values were two-sided. Statistical significance was defined as $p<0.05$.

Variable	Non-Recurrence CAD(n=5222)	Recurrence CAD(n=2361)	P-Value
Age yrs.(range)	70(18-88)	69(31-88)	0.7364
Gender			0.2117
Male (%)	3582(69)	1632(70)	
Female (%)	1640(31)	729(30)	
Platelet (K/uL)	202.78(160.42-255.67)	205.94(160.25-260.12)	0.2664
WBC (K/uL)	10.32(8.43-12.69)	9.87(7.8-12.2)	<0.0001
Hematocrit (%)	28(0-32.8)	25.4(0-32)	<0.0001
Total calcium (mg/dL)	8.44(8.02-8.81)	8.47(8.09-8.84)	0.0014
Chloride (mEq/L)	104.12(101.78-106.25)	103.53(100.8-105.87)	<0.0001
LDL (mg/dL)	10.53(±0.27)	8.36(±0.36)	<0.0001
CK (IU/L)	65.17(0-260.75)	63.6(0-192)	0.5732
Creatinine (mg/dL)	1(0.8-1.34)	1.1(0.85-1.64)	<0.0001
Uric acid (mg/dL)	0.31(±0.02)	0.31(±0.03)	0.8964
INR	1.25(1.13-1.44)	1.26(1.14-1.5)	0.0021
HbA1c (%)	1.75(±0.04)	1.34(±0.06)	<0.0001
RBC (m/uL)	3.52(3.24-3.9)	3.48(3.21-3.86)	0.0005

Model's performance

The AUCs discriminatory abilities of all three models for the prediction of CAD recurrence within 6 months is shown in Figure 2 (A) AUROC and (B) AUPR. The performance of each model was

generated from the testing set. The XGBoost, RF and LR had AUCs of 0.81, 0.83 and 0.63 respectively. The AUC of XGBoost model was 0.74, recall 0.76, specificity 0.76 and precision 0.75. The AUC of RF model was 0.76, recall 0.80, specificity 0.73, precision 0.78 and AUC of LR is 0.60, recall 0.61, specificity 0.58 and precision 0.60 (Table 2).

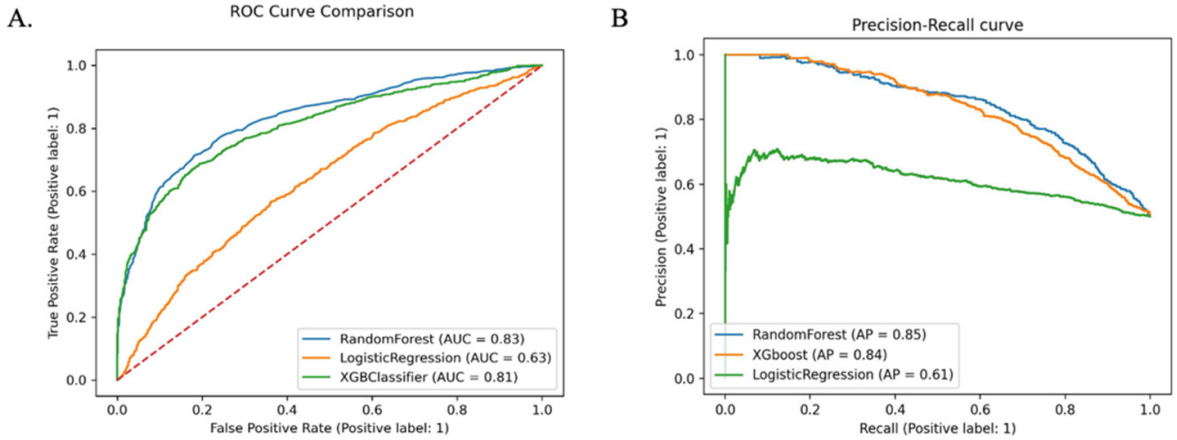


Figure 2: Receiver operator characteristic curves and precision-recall curve. (A) Receiver operating characteristic curve. (B) Precision-recall curve. XGBoost extreme gradient boosting. ROC operating characteristic curve. AUC area under the curve. AUPRC area under the precision-recall curve.

Table 2: Model performance in the testing dataset.

CI confidence interval; AUROC area under curve under the receiver operating characteristic curve; XG Boost extreme gradient boosting, LR logistic regression, RF random forest.

Model	Accuracy (95%CI)	Precision (95%CI)	F1score (95%CI)	AUROC (95%CI)	Recall (95%CI)	Specificity (95%CI)	AUPRC
XG Boost	0.74(0.65–0.85)	0.75(0.67–0.82)	0.73(0.64–0.81)	0.81(0.65–0.81)	0.76(0.65–0.81)	0.72(0.75–0.79)	0.84
RF	0.76(0.74–0.78)	0.78(0.71–0.81)	0.76(0.69–0.80)	0.83(0.69–0.84)	0.80(0.69–0.81)	0.73(0.71–0.79)	0.85
LR	0.60(0.58–0.62)	0.60(0.58–0.62)	0.60(0.58–0.62)	0.63(0.56–0.65)	0.61(0.58–0.62)	0.58(0.60–0.65)	0.61

Feature importance

The feature importance was assessed using the XGBoost model and RF model. In the XGBoost model, the top six (6) important variables in predicting the recurrence of CAD were platelet, WBC, RBC, INR, chloride, and creatinine. In the RF model, the top six (6) are creatinine, platelet, WBC, chloride, RBC, and total calcium (Figure 3). The top 6 features in XGBoost model were similar to those in RF model.

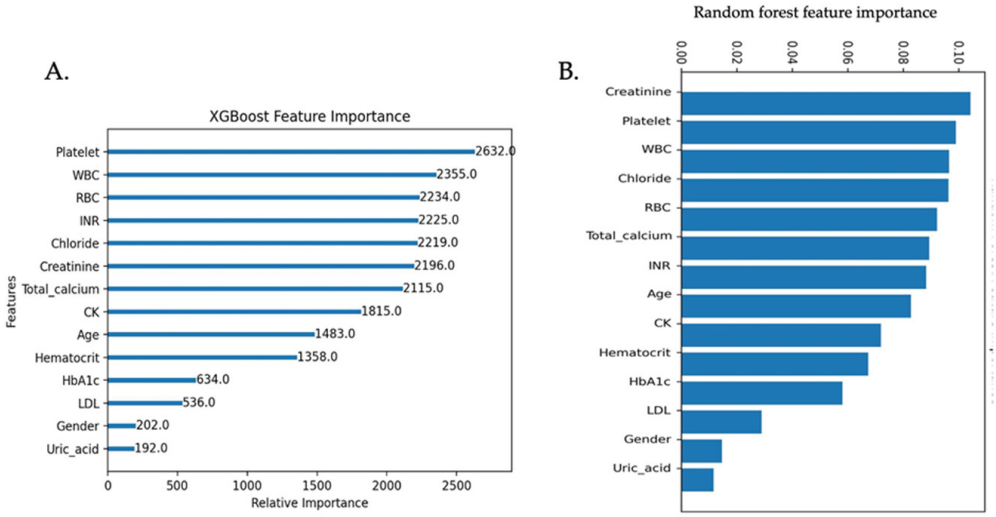


Figure 3: A. Features importance in a classifier based on XGBoost. B. Features importance in a classifier based on the Random Forest model. WBC white blood cell, RBC red blood cell, INR international normalized ratio, CK creatinine kinase, HbA1c hemoglobin A1C, LDL low-density lipoprotein.

Multivariate logistic regression analysis of factors for CAD reoccurrence

Multivariate logistic regression analysis identified that gender, platelet count, total calcium, creatinine, and INR were associated with a 1-fold increase in CAD recurrence when it increased by one unit (OR:1.082, OR: 1.100, OR: 1.045, OR:1.116, OR: 1.096) respectively (Table 3).

Table 3: Multivariate analysis of factors for CAD recurrence.

WBC: White Blood Cell, RBC: Red Blood Cell, INR: International Normalized Ratio, CK: Creatinine Kinase, HbA1c: Hemoglobin A1C, LDL: Low-Density Lipoprotein. mg/dl milligrams per deciliter, IU/L: International units per litre, cm: Centimeter (cm), yrs: Years, S.D: Standard Deviation, K/uL: Thousand per microliter, m/uL million per microliter, % percentage, mEq/L milliequivalents per liter. CI, confidence interval.

Variable	Odd Ratio	(95% CI)	P-value
Age, yr	0.998	(0.994-1.002)	0.33
Gender	1.082	(0.970-1.208)	0.156
Platelet (K/uL)	1.1	(1.000-1.001)	0.129
WBC (K/uL)	0.987	(0.976-0.998)	0.026
Hematocrit (%)	0.992	(0.988-0.995)	<0.001
Total calcium (mg/dL)	1.045	(1.020-1.070)	<0.001
Chloride (mEq/L)	0.988	(0.980-0.995)	0.001
LDL (mg/dL)	0.997	(0.994-0.999)	0.018
CK (IU/L)	1	(1.000-0.999)	0.022
Creatinine (mg/dL)	1.116	(1.071-1.163)	<0.001
Uric acid (mg/dL)	0.983	(0.953-1.014)	0.282
INR	1.096	(1.019-1.180)	0.014
HbA1c (%)	0.964	(0.946-0.983)	<0.001
RBC (m/uL)	0.881	(0.801-0.968)	0.009

Discussion

Main findings

To the best of our knowledge, this study represents the first investigation into the recurrence of CAD within a 6-month post-treatment period. The study's findings highlight the significant clinical role of Machine Learning (ML) in evaluating the prognostic risk of CAD recurrence. The primary focus of this study is on assessing the predictive power of laboratory measurements and patients demographic information in CAD recurrence after treatment using machine learning algorithms. This approach aims to identify potential biomarkers or indicators of CAD risk. By examining lab data, the study explores the association between specific lab measurements and the occurrence or recurrence of CAD, providing insights into the independent predictive role of these lab variables, regardless of other known risk factors. The Random Forest (RF) algorithm showcased exceptional predictive performance in our study, yielding an impressive Area Under the Curve (AUC) value of 0.83 and accuracy of 0.76. These findings surpassed the results obtained from the integrated tree-based XGBoost algorithm and logistic regression. A study conducted by Aravind et al. [35] employed clinical data to construct predictive models using Machine Learning (ML) algorithms, aiming to aid clinicians in the timely detection of Coronary Artery Disease (CAD). Their RF model achieved a remarkable accuracy of 0.87. Similarly, another investigation, led by Jinwan et al. [36] aimed to develop prediction models using ML algorithms to anticipate the risk of major adverse cardiovascular events within 6 months post-coronary revascularization. Their RF model, following oversampling with SMOTE, demonstrated a performance with an accuracy of 0.75.

In the absence of traditional CAD risk factors such as alcohol consumption, smoking, diabetes, obesity, hypertension, and history of heart disease, our model successfully predicted CAD recurrence within the 6-month post-treatment period. The Random Forest model can serve as a decision support tool for clinicians, complementing their expertise and facilitating more informed decisions regarding treatment plans, interventions, and lifestyle modifications for patients with suspected or diagnosed CAD. Furthermore, the model's performance may contribute to the development or refinement of clinical guidelines for CAD management, promoting standardized and evidence-based approaches in clinical practice.

In terms of age distribution, our study observed similarities between CAD recurrent and non-recurrent patients. However, CAD recurrence exponentially increased with advancing age, consistent with previous studies [37,38]. Our findings also align with epidemiological studies indicating a higher incidence of CAD in males compared to females [39]. However, a study by Miller et al. [40] reported no gender differences in CAD patients and those with CAD recurrence. Despite advances in health technology for diagnosing and treating cardiovascular diseases, CAD remains the leading cause of morbidity and mortality for both men and women worldwide [41]. A clinical study conducted in the Netherlands over a 16-year period, evaluating 1,894 patients with coronary angiography, found no gender difference in the extent of coronary lesions observed [42]. Efforts to reduce these rates should focus on improving the implementation of guideline-directed recommendations after CAD diagnosis, including treatment protocols. Although medications such as statins, aspirins, ACE inhibitors, and antiplatelet and anticoagulant combination therapy reduce the risk of recurrent events [13], the adherence to these treatment regimens remains insufficient, and many high-risk patients do not continue with guideline-recommended treatments [43].

In the multivariate analysis, we identified gender, platelet count, total calcium, creatinine, and INR as independent predictors closely associated with CAD recurrence. These features indicate an increased risk of CAD recurrence after treatment. Feature importance was assessed using the f-score from both the XGBoost and Random Forest models. The top six most important features identified were platelet count, White Blood Cell Count (WBC), Red Blood Cell Count (RBC), INR, chloride, and creatinine. Feature

importance was assessed using an f-score from the XGBoost model and random forest model. This metric measures the number of times a feature is used to split the data across all trees. From the XGBoost model, and RF model, the top 6 most important features are platelet, WBC, RBC, INR, chloride, and creatinine.

Clinical implications

Machine learning models offer a concrete and empirical screening method to estimate a patient's potential to develop specific diseases like CAD. While clinicians traditionally rely on personal judgment, which can be subjective and grounded in their clinical expertise, machine learning provides an objective approach to CAD diagnosis. These models can seamlessly integrate into everyday clinical practice without requiring expensive medical equipment. Consequently, clinicians can efficiently assess a patient's CAD risk promptly, whether in real-time or over a duration, without adding to their workload. This significance is evident not only in the rapidly evolving healthcare systems of developed countries but also in the resource-constrained environments of developing nations.

By assessing specific lab biomarkers and integrating them with Machine Learning (ML), healthcare providers can gain a deeper understanding of disease progression. In this present study, our ML models, which are developed based on the disease biomarkers, have proven to provide valuable insights into the treatment response of CAD and its probable recurrence. The use of laboratory data allows for objective and quantitative measurements, providing a standardized and reliable assessment of patients' conditions. This approach enhances the accuracy and precision of monitoring, enabling early detection of recurrent CAD and prompt intervention. When these models are integrated into clinical practice, healthcare professionals could optimize treatment strategies, tailor interventions based on individual patient profiles, and improve overall patient outcomes. Therefore, the inclusion of laboratory data and ML in this study enhances its clinical relevance and usefulness in informing patient care decisions. The outcome of our study underscore the promise of future investigations employing these algorithms to further refine CAD diagnosis accuracy. The pursuit of heightened precision in disease prediction and prevention demands the integration of sophisticated tools as we confront the challenges posed by CAD.

Limitations

The study had limitations, including its retrospective nature and reliance on existing medical records, which could lead to incomplete or missing data. Although efforts were made to clean the data for the present study, inaccurate or incomplete data could affect the reliability of the findings. Additionally, the 6-month follow-up duration may not capture longer-term CAD recurrence patterns, warranting longer follow-up durations for a more comprehensive understanding of disease progression. While the study may establish associations between laboratory indices and CAD recurrence, it cannot establish causality. Other factors not considered in the study may contribute to CAD recurrence. Considering CAD's complexity, incorporating additional risk factors into predictive models can enhance accuracy. Addressing these limitations in future research can enhance understanding of CAD recurrence and the utility of laboratory indices in predicting and monitoring disease progression.

Future work

Expanding upon the insights garnered from our current investigation, promising avenues for future research emerge, poised to further enrich our comprehension of CAD. These potential future studies not only broaden the scope of our findings but also carry the potential to address previously unexplored facets and amplify the practical implications of our research. In forthcoming endeavors, we aspire to delve into the realm of Clinical Scoring Systems Comparison. This vital endeavor entails conducting a comprehensive assessment that scrutinizes the performance of machine learning algorithms against established clinical scoring systems, such as the widely recognized Framingham Risk Score. This deliberate evaluation seeks to unveil whether machine learning possesses the potential to surpass conventional scoring systems in terms of predictive accuracy for CAD recurrence. Additionally, we aim to undertake Extended Follow-Up Analysis. This trajectory of investigation extends beyond the immediate 6-month post-treatment period, immersing itself in prolonged observation windows spanning 1 to 2 years. This exploration seeks to illuminate the sustained predictive prowess of machine learning algorithms across these extended durations. These longer timeframes could potentially unravel the temporal dynamics governing CAD recurrence risk. It involves evaluating whether the predictive insights gleaned within the initial 6 months remain steadfast over these extended periods, thereby yielding invaluable insights into the enduring effectiveness of predictive models.

Conclusion

The RF model outperforms the XGBoost and LR models in predicting CAD recurrence within 6 months after treatment. The study indicates a potential association between laboratory indices (platelet, WBC, RBC, INR, chloride, total calcium, and creatinine) and CAD recurrence. These findings contribute to filling gaps in knowledge, inspiring further research, and providing guidance for future investigations into preventive strategies and treatment approaches regarding CAD.

Acknowledgment

The data was acquired from Medical Information Mart for Intensive Care (MIMIC-III 1.4).

References

1. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Pankaj Arora, et al. (2023) Heart disease and stroke statistics-2023 update: A report from the American Heart Association. Circulation 147(8): e93-e621.
2. Eberhard Standl, Kamlesh Khunti, Hansen TB, Oliver Schnell (2019) The global epidemics of diabetes in the 21st century: Current situation and perspectives. Eur J Prev Cardiol 26(2_suppl): 7-14.
3. Rupert Bauersachs, Uwe Zeymer, Jean-Baptiste Brière, Caroline Marre, Kevin Bowrin, et al. (2019) Burden of coronary artery disease and peripheral artery disease: A literature review. Cardiovasc Ther 2019: 8295054.

4. Ahmed Jamal, Elyse Phillips, Gentzke AS, Homa DM, Babb SD, et al. (2018) Current cigarette smoking among adults-United States, 2016. MMWR Morb Mortal Wkly Rep 67(2): 53-59.

5. Hugues Sampasa-Kanyinga, Lewis RF (2015) Frequent use of social networking sites is associated with poor psychological functioning among children and adolescents. Cyberpsychol Behav Soc Netw 18(7): 380-385.

6. Kuulasmaa K, Pedoe KT, Dobson A, Fortmann S, Sans S, et al. (2000) Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. The lancet 355(9205): 675-687.

7. McGovern PG, Pankow JS, Shahar E, Doliszny KM, Folsom AR, et al. (1996) Recent trends in acute coronary heart disease-mortality, morbidity, medical care, and risk factors. The Minnesota heart survey investigators. N Engl J Med 334(14): 884-890.

8. Murray CJL, Theo Vos, Rafael Lozano, Mohsen Naghavi, Flaxman AD, et al. (2012) Disability-Adjusted Life Years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. The Lancet 380(9859): 2197-2223.

9. Moran AE, Forouzanfar MH, Roth GA, Mensah GA, Majid Ezzati, et al. (2014) Temporal trends in ischemic heart disease mortality in 21 world regions, 1980 to 2010: The global burden of disease 2010 study. Circulation 129(14): 1483-1492.

10. Weintraub NL, Collins SP, Pang PS, Levy PD, Anderson AS, et al. (2010) Acute heart failure syndromes: Emergency department presentation, treatment, and disposition: Current approaches and future aims: A scientific statement from the American Heart Association. Circulation 122(19): 1975-1996.

11. Simon Barquera, Pedroza-Tobías A, Catalina Medina, Hernández-Barrera L, Bibbins-Domingo K, et al. (2015) Global overview of the epidemiology of atherosclerotic cardiovascular disease. Arch Med Res 46(5): 328-338.

12. Roser M, Ritchie H, Spooner F (2020) Burden of disease. Our World in Data.

13. Waljee AK, Higgins PD (2010) Machine learning in medicine: A primer for physicians. Am J Gastroenterol 105(6): 1224-1226.

14. Hajar R (2017) Risk factors for coronary artery disease: Historical perspectives. Heart Views 18(3): 109-114.

15. Sajda P (2006) Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng 8: 537-565.

16. Foster KR, Koprowski R, Skufca JD (2014) Machine learning, medical diagnosis, and biomedical engineering research-commentary. Biomed Eng Online 13: 94.

17. Pooja Rani, Rajneesh Kumar, Ahmed NMS, Anurag Jain (2021) A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments 7(3): 263-275.

18. Henrietta Forssen, Riyaz Patel, Natalie Fitzpatrick, Aroon Hingorani, Adam Timmis, et al. (2017) Evaluation of machine learning methods to predict coronary artery disease using metabolomic data. Stud Health Technol Inform 235: 111-115.

19. Chowdary GJ, Suganya G, Premalatha M (2020) Effective prediction of cardiovascular disease using cluster of machine learning algorithms. Journal of Critical Reviews 7(19): 1865-1875.

20. Paynter NP, Raji Balasubramanian, Franco Giulianini, Wang DD, Tinker LF, et al. (2018) Metabolic predictors of incident coronary heart disease in women. Circulation 137(8): 841-853.

21. Sabarish K, Parvati T (2021) An experimental investigation on L9 orthogonal array with various concrete materials. Materials Today: Proceedings 37(2): 3045-3050.

22. Nasser IM, Abu Naser SS (2019) Lung cancer detection using artificial neural network. International Journal of Engineering and Information Systems (IJEAIS) 3(3): 17-23.

23. El_Jerjawi NS, Abu Naser SS (2018) Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology 121: 55-64.

24. Johnson AEW, Pollard TJ, Lu Shen, Lehman LH, Mengling Feng, et al. (2016) MIMIC-III, a freely accessible critical care database. Sci data 3:160035.

25. Chawla, NV (2010) Data mining for imbalanced datasets: An overview. Data Mining and Knowledge Discovery Handbook 853-867.

26. Harrell FE (2001) Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. In: (2nd edn), Springer Publishers, USA, pp. 582.

27. Oommen T, Baise LG, Vogel RM (2011) Sampling bias and class imbalance in maximum-likelihood logistic regression. Mathematical Geosciences 43: 99-120.

28. Zheng H, J Yuan, L Chen (2017) Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. Energies 10(8): 1168.

29. Chen X, Wang ZX, Pan XM (2019) HIV-1 tropism prediction by the XGboost and HMM methods. Scientific reports 9(1): 9997.

30. Titapiccolo JI, Manuela Ferrario, Sergio Cerutti, Carlo Barbieri, Flavio Mari, et al. (2013) Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. Expert Systems with Applications 40(11): 4679-4686.

31. Breiman L (2001) Random forests. Machine Learning 45: 5-32.

32. Azamossadat Hosseini, Eshraghi MA, Tania Taami, Hamidreza Sadeghsalehi, Zahra Hoseinzadeh, et al. (2023) A mobile application based on efficient lightweight CNN model for classification of B-ALL cancer from non-cancerous cells: A design and implementation study. Informatics in Medicine Unlocked 39: 101244.

33. Powers D (2011) Evaluation: From precision, recall and F-Factor to ROC, informedness, markedness & correlation. Mach Learn Technol 2: 37-63.

34. Ali Garavand, Cirruse Salehnasab, Ali Behmanesh, Nasim Aslani, Zadeh AH, et al. (2022) Efficient model for coronary artery disease diagnosis: A comparative study of several machine learning algorithms. J Healthc Eng 2022: 5359540.

35. Akella A, Akella S (2021) Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution. Future sci OA 7(6): FSO698.

36. Jinwan Wang, Shuai Wang, Xuefang ZM, Tao Yang, Qingfeng Yin, et al. (2022) Risk prediction of major adverse cardiovascular events occurrence within 6 months after coronary revascularization: Machine learning study. JMIR Med Inform 10(4): e33395.

37. Giorda CB, Angelo Avogaro, Marina Maggini, Flavia Lombardo, Edoardo Mannucci, et al. (2008) Recurrence of cardiovascular events in patients with type 2 diabetes: Epidemiology and risk factors. Diabetes Care 31(11): 2154-2159.

38. Cubbon RM, Afroze Abbas, Wheatcroft SB, Niamh Kilcullen, Raj Das, et al. (2008) Diabetes mellitus and mortality after acute coronary syndrome as a first or recurrent cardiovascular event. Plos One 3(10): e3483.

39. Greenland P, Reiss HR, Goldbourt U, Behar S (1991) In-hospital and 1-year mortality in 1,524 women after myocardial infarction. Comparison with 4,315 men. Circulation 83(2): 484-491.

40. Miller TD, Roger VL, Hodge DO, Hopfenspirger MR, Bailey KR, et al. (2001) Gender differences and temporal trends in clinical characteristics, stress test results and use of invasive procedures in patients undergoing evaluation for coronary artery disease. J Am Coll Cardiol 38(3): 690-697.

41. Steingart RM, Packer M, Hamm P, Coglianese ME, Gersh B, et al. (1991) Sex differences in the management of coronary artery disease. Survival and ventricular enlargement investigators. N Engl J Med 325(4): 226-230.

42. Lennep JER, Zwinderman AH, Lennep HWR, Westerveld HE, Plokker HW, et al. (2000) Gender differences in diagnosis and treatment of coronary artery disease from 1981 to 1997. No evidence for the Yentl syndrome. Eur heart J 21(11): 911-918.

43. Peters SAE, Colantonio LD, Yuling Dai, Hong Zhao, Vera Bittner, et al. (2021) Trends in recurrent coronary heart disease after myocardial infarction among US women and men between 2008 and 2017. Circulation 143(7): 650-660.