Machine Learning-Based Classification of Coronary Artery Disease Using Complete Blood Count and Basic Clinical Variables

Inzimam Ul Haq[1], Syed Asim Jalal[1]*, Ayesha Khattak[2], Muhammad Sadeequllah[1], Ibrar Ahmed[1], Saif ur Rehman[1]
[1]Department of Computer Science, University of Peshawar
[2]School of Computer Science and IT, Institute of Management Sciences, Peshawar, Pakistan
***Correspondence**: sa.jalal@uop.edu.pk

Citation | Haq. I. U, Jalal. S. A, Khattak. A, Sadeequllah. M, Ahmed. I, Rehman. S. U, "Machine Learning-Based Classification of Coronary Artery Disease Using Complete Blood Count and Basic Clinical Variables", IJIST, Vol. 8 Issue. 10 pp 4108-4121, November 2026
Received | Oct 26,2026 **Revised |** Nov 08, 2026 **Accepted |** Nov 10, 2026 **Published |** Nov 12, 2026.

This retrospective study examined whether a restricted set of routinely collected complete blood count (CBC) variables, combined with basic demographic and clinical factors, can distinguish records labelled as coronary artery disease (CAD) from non-CAD records using conventional machine-learning classifiers. Anonymized hospital records collected from a local teaching hospital between November 2017 and November 2022 were retrospectively processed. After complete-case filtering, 3,301 records remained, including 1,607 CAD and 1,694 non-CAD records. Twelve predictors were used: age, gender, hypertension, diabetes, hemoglobin, total leukocyte count, neutrophils, lymphocytes, monocytes, eosinophils, hematocrit, and platelets. Six classifiers were evaluated using an 80/20 train-test split. On the 661-record held-out test set, Random Forest achieved the highest accuracy (81.24%) and F1-score (81.04%), Logistic Regression achieved the highest precision (84.72%), and Naïve Bayes achieved the highest recall/sensitivity (79.65%). These findings suggest that a compact set of CBC and basic clinical variables contains useful information for CAD classification in this retrospective cohort. However, the single-center design, complete-case selection, lack of feature-set ablation, external validation, and calibration assessment limit clinical interpretation. The findings should therefore be considered exploratory rather than evidence of a validated diagnostic tool.

Keywords: Coronary Artery Disease Classification; Machine Learning; Complete Blood Count; Routine Clinical Data

Introduction:

Cardiovascular diseases (CVDs) are among the major causes of death worldwide and continue to create a considerable burden on healthcare systems [1]. Early identification of people at risk of cardiovascular disease can support timely medical assessment and treatment. Due to the increasing availability of electronic health records and clinical data, machine learning techniques are now being widely studied for supporting cardiovascular disease prediction. Different machine learning models have been developed to analyze patient information and assist medical professionals in identifying patients who may require further investigation [2]. Machine learning-based prediction of heart disease usually depends on different types of patient information. These may include demographic information, medical history, family history, cardiovascular risk factors, laboratory investigations, electrocardiogram (ECG), echocardiography, and other clinical examination results [3]. The availability of these variables can improve model performance, but it can also make the prediction process dependent on a large number of clinical measurements. In practical settings, all these investigations may not always be available at the same time, particularly during an initial patient assessment.

Among cardiovascular diseases, coronary artery disease (CAD), also known as coronary heart disease, is one of the most common forms of heart disease [4]. Plaque buildup inside the coronary arteries is the usual cause of CAD. Blood flow to the heart is reduced as plaque accumulates, and severe cases may involve partial or complete arterial blockage [5]. CAD is assessed using procedures and tests such as ECG, echocardiography, exercise stress testing, and coronary angiography. These methods provide important clinical information, but some of them require specialized equipment, trained medical staff, additional time, or higher cost.

For this reason, a considerable amount of research has investigated the use of machine learning for the prediction and classification of CAD. Many studies have used publicly available heart disease datasets for developing and comparing machine learning models. Two commonly used examples are the Z-Alizadeh Sani dataset [6] and the UCI Heart Disease dataset [7]. These datasets contain several types of variables, including demographic, clinical, laboratory, ECG, and other examination-related information. Previous studies have used different subsets of these variables and have reported varying levels of prediction performance.

Although the use of a larger number of clinical variables can help in building predictive models, it also means that similar information may be required when the model is applied to new patients. This raises a different question: how much useful discrimination remains when the input feature space is deliberately restricted to routine hematological and basic clinical variables? Complete Blood Count (CBC), also referred to as Full Blood Count (FBC), is a commonly performed laboratory test and provides several hematological measurements. Many hospitals routinely have these measurements available, without the need for specialized cardiovascular investigations.

Since 2022, studies have used hematological and other routinely available laboratory information for cardiovascular event prediction, CBC-derived risk stratification, early-stage CAD, premature CAD, and broader cardiovascular disease classification [8] [9] [10] [11] [12] [13] [14]. This work extends the use of such information across cardiovascular prediction and classification tasks and supports continued interest in accessible blood-based information. Its outcomes, feature sets, and validation designs differ, however, from those of the restricted CAD classification problem studied here.

The retrospective analysis used anonymized records containing relevant clinical information and CBC/FBC test results from a local teaching hospital in Peshawar, Pakistan. After cleaning and exclusion of incomplete records, the dataset comprised 1,607 CAD-labelled

and 1,694 non-CAD records, giving a total of 3,301. To examine how well CAD and non-CAD cases could be distinguished with this restricted set of routinely available variables, six commonly used machine learning algorithms were evaluated using CBC-related variables combined with basic demographic and clinical predictors such as age, gender, hypertension, and diabetes.

Research Objectives:

To evaluate whether a restricted combination of CBC parameters and basic demographic/clinical variables can discriminate between CAD and non-CAD records.

To compare the performance of six conventional machine-learning classifiers on this combined feature set using accuracy, precision, recall/sensitivity, specificity, F1-score, and confusion matrices.

Research Novelty:

The novelty of this study lies not in proposing a new machine-learning algorithm, but in evaluating CAD classification using a deliberately restricted combination of routinely available CBC parameters and basic clinical variables from real hospital records. Unlike studies using broader clinical, biochemical, ECG, echocardiographic, or imaging information, the present study examines the classification performance retained under this limited information set.

Main Contributions:

A retrospective hospital dataset of 3,301 records was prepared using data collected from a real clinical setting with a restricted set of routinely available variables.

A constrained-feature evaluation was conducted to examine CAD discrimination when the available information was limited to 12 routine hematological and basic clinical variables.

Several classifiers with different modelling assumptions produced similar overall performance, providing exploratory evidence that the discriminatory signal in the combined restricted feature set was not specific to one algorithm.

The remainder of this paper is organized as follows. Section 2 reviews related studies on machine learning-based prediction of coronary artery and heart disease. Section 3 describes the dataset, preprocessing steps, and machine learning methods used in the study. Section 4 presents the experimental results, followed by the discussion of the findings and their limitations. Finally, the last section concludes the paper and presents directions for future work.

Related work:

Most existing machine-learning studies on heart disease and coronary artery disease (CAD) have used public benchmark datasets to compare the performance of classification algorithms. These datasets include the Z-Alizadeh Sani and UCI Heart Disease datasets, which are used with Support Vector Machines, K-Nearest Neighbors, Random Forests, Decision Trees, Naïve Bayes, Logistic Regression, neural networks, along with various feature-selection and optimization techniques [15][16][17][18][19][20][21] [22][23]. The results of these studies generally show that machine learning classifiers can detect useful patterns in clinical data and can predict diseases with reasonable performance. The reported performance values, however, vary between studies. Such variation is typically expected of differences in feature subsets, preprocessing methods, training and testing procedures, and validation schemes adopted by different studies. Due to these differences, the accuracy reported by these research studies cannot be directly compared with each other.

The datasets used in many of these studies have some limitations. The Z-Alizadeh Sani dataset has only 303 patient records, although it contains a relatively large number of features based on clinical, laboratory, ECG, and echocardiographic variables [6]. The Cleveland Heart Disease dataset is also small and has 303 records [7]. Both of these datasets have been useful

for testing and comparing algorithm performance, but these feature sets are not fully comparable to the variations found in typical routine hospital patients' data. Furthermore, models trained on small benchmark datasets may perform differently when applied to data from another region, hospital, or population.

The Framingham Heart Study is a large and important source of cardiovascular data [24]. It has been used in many machine-learning studies. The purpose and structure of the Framingham data are, however, different from those of a binary classification task of presence or absence of CVD. Cardiovascular risk, heart disease, and confirmed CAD are different prediction problems. Such differences in the target variable and the number of classes also affect model performance. Therefore, prediction performance results from different datasets should be interpreted appropriately.

Another important aspect of the study area is the use of real-world hospital and electronic health record data. Li et al. used over 100,000 electronic medical records in their cardiovascular disease prediction study [25]. Some studies [17][26][27] developed and used their own hospital based datasets in Algeria, Nigeria, and China. Naushad et al. similarly examined machine-learning for CAD risk prediction using a clinical cohort and reported useful prediction performance [28]. These studies are important because their data come from actual healthcare settings rather than public benchmark datasets. Such work gives stronger evidence about how machine-learning models may behave when patient data come from real clinical settings.

The use of local and regional data is also important because the characteristics of patients can differ across populations. [29] evaluated machine-learning algorithms for cardiovascular disease prediction using patient data collected from two teaching hospitals in Khyber Pakhtunkhwa, Pakistan. Both studies draw on data from healthcare settings in the same wider region, making that work relevant to the present study.

The inputs required by a model also deserve attention, both in number and type. Age, symptoms, blood pressure, cholesterol, ECG findings, exercise-test information, imaging results, and other clinical measurements are often combined in heart disease prediction studies [15], [19], [25], [26], [27]. A larger set of variables can improve prediction performance, but the resulting dependence on information that is not always available can limit a model's usefulness. This may be the case when a model requires ECG, echocardiography, imaging, and several laboratory tests in a setting where only basic clinical information is available. These requirements shift the research question from maximizing performance using all available information to examining how much useful discrimination remains when the feature set is deliberately restricted.

This question makes routine laboratory tests of particular interest. Among the most commonly performed is the Complete Blood Count (CBC), also known as the Full Blood Count (FBC). Its measurements include hemoglobin, white blood cell count, hematocrit, platelets, neutrophils, lymphocytes, monocytes, and other hematological parameters. These variables reflect different physiological processes, some of which may be related to CVD. For example, white blood cells and neutrophil-related measures can provide information about inflammatory activity, while hemoglobin and hematocrit are related to the oxygen-carrying capacity of blood. Platelet measurements are also relevant because platelets are involved in thrombosis. Routine hematological information may therefore carry useful signals even when collected for purposes other than cardiovascular diagnosis.

Evidence for this possibility comes from studies of cardiovascular risk assessment. [8] used hematological predictors with artificial intelligence methods, and their findings suggest that routinely available blood measurements can contribute useful information to risk assessment. For early risk stratification of acute myocardial infarction, [9] developed a score from complete blood count information. That outcome differs from CAD classification, but

the study demonstrates the potential value of existing CBC results in identifying cardiovascular risk.

Hematological and laboratory markers have been examined more directly in recent machine-learning studies. With 3,316 patients and clinical and routine laboratory biomarkers, the early-stage CAD prediction study by [10] is particularly relevant to the present work in both sample size and use of routinely collected variables. [11] applied machine learning and explainable AI to acute myocardial infarction classification using hematological markers; white blood cell and neutrophil-related information were among the useful predictors they identified. For premature coronary artery disease, [12] studied machine-learning models using routine hematological measurements, including white blood cell count, hemoglobin, hematocrit, neutrophils, and platelets.

Taken together, these findings suggest that routine hematological variables may be useful for CAD-related classification, particularly in combination with other patient characteristics.

A closer example is the recent retrospective study of 896 hospital patients by [13]. Across five machine-learning classifiers, the authors compared a baseline clinical feature set with an enhanced set that added routine hematology and biochemistry results. Model performance improved with the blood-test variables, and the best XGBoost model achieved 91.62% accuracy. SHAP analysis was used to examine how individual laboratory markers contributed. The authors included routine hematological variables such as platelet count, hemoglobin, white blood cell count, red blood cell count, and hematocrit together with biochemical measurements. This study provides recent evidence that routinely available blood-test information can add useful information to cardiovascular disease classification.

The value of routine blood measurements should nevertheless be interpreted carefully. CBC variables are affected by many conditions and are not specific markers of CAD. Changes in leukocytes, hemoglobin, platelets, or hematocrit can occur due to infection, inflammation, anemia, medication, and several other medical conditions. Therefore, the purpose of using these measurements in a machine-learning model is not to claim that a CBC test can diagnose CAD by itself. A more reasonable question is whether patterns contained in these routinely available measurements can contribute useful information when they are combined with basic demographic and clinical risk factors.

There has also been increasing attention to the methodological quality of cardiovascular machine-learning studies. A recent systematic review and meta-analysis by [14] examined machine-learning models developed from electronic health record data for cardiovascular risk prediction. The review reported promising results with some insights. It highlighted differences among different studies in the datasets used, models developed, their validation procedures, and reporting. It also noted limited external validations. These issues are important when understanding results of the studies that report very high classification accuracy. These issues are even more important when the studies use small or commonly reused datasets.

The main contribution of our study is not a comparison to identify the best machine-learning classifier. Instead, the classifiers are used as different modelling approaches to test the same constrained-information question.

The study asks whether a restricted combined feature set can provide useful discrimination between CAD and non-CAD records. This question is examined using 3,301 hospital records and 12 predictors consisting of CBC parameters, age, gender, hypertension, and diabetes. The inputs excluded specialized cardiovascular information, including ECG, echocardiography, and imaging.

The remaining question is how well routine hospital records can support CAD classification when only CBC parameters and a few basic clinical variables are used. Recent

blood-based studies have addressed cardiovascular risk, acute myocardial infarction, premature CAD, and broader cardiovascular disease classification, often drawing on different or more extensive clinical and laboratory inputs [8][9][10][11][12][13][14]. Substantial differences in cardiovascular outcomes, feature sets, patient populations, and validation designs prevent direct comparison of previous studies with the present results. This study does not establish how much CBC contributes beyond the basic clinical factors.

Materials and Methods:

Study Design and Data Source:

Anonymized patient records collected between November 2017 and November 2022 at a local teaching hospital in Peshawar, Pakistan, were used for this retrospective study. Within the hospital database, separate MySQL tables held patient visits, laboratory reports, and diagnosis information. The study focused on records having the required clinical information together with Full Blood Count (FBC), also referred to as Complete Blood Count (CBC), test results. The study design workflow is shown in Figure 1.

Record Selection and Data Preprocessing:

Approximately 22,400 candidate records were retrieved from the hospital database. Records were excluded when CBC reports were incomplete, required demographic or basic clinical values were missing, or unusually high CBC measurements were removed after consultation with a specialist hematologist. No interpolation or imputation was used. The final dataset contained 3,301 records (1,607 CAD-labelled and 1,694 non-CAD), leaving approximately 19,100 candidate records not retained.

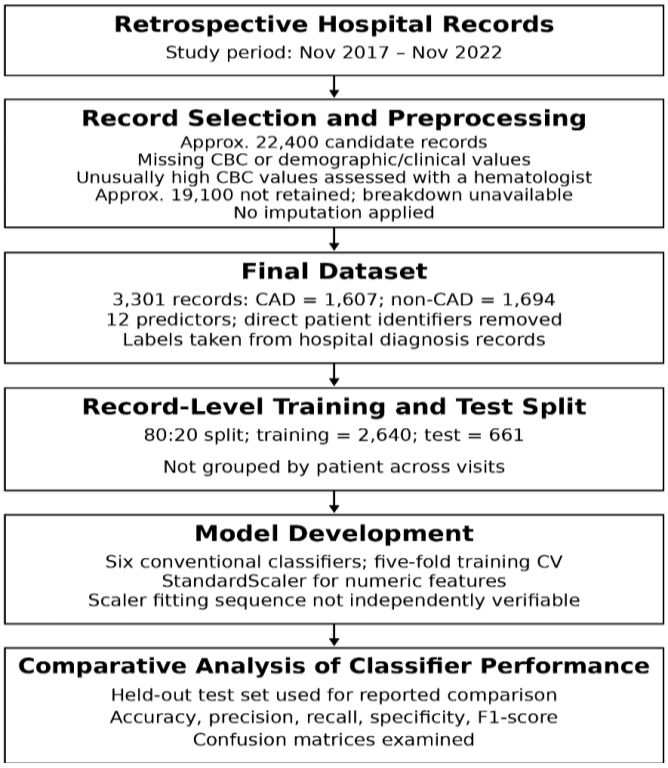


Figure 1. Workflow of record selection, record-level splitting, model development, and final classifier comparison.

Predictor and Outcome Variables:

Of the 12 predictor features, eight were CBC measurements: hemoglobin, total leukocyte count (TLC), neutrophils, lymphocytes, monocytes, eosinophils, hematocrit (HCT), and platelet count. Age, gender, hypertension, and diabetes provided the four basic demographic or clinical predictors. There was one binary target variable, CAD status, with CAD coded as 1 and non-CAD as 0. Table 1 lists the predictors and target variable.

Table 1. Predictor and outcome variables used in the study.

Variable	Description / Coding	Unit
Age	Patient age	Years
Gender	Male = 0; Female = 1	Binary
Hypertension	Yes = 1; No = 0	Binary
Diabetes	Yes = 1; No = 0	Binary
Hemoglobin	Hemoglobin level	g/dL
TLC	Total leukocyte count	$10^3/\mu L$
Neutrophils	Neutrophil percentage	%
Lymphocytes	Lymphocyte percentage	%
Monocytes	Monocyte percentage	%
Eosinophils	Eosinophil percentage	%
HCT	Hematocrit	%
Platelets	Platelet count	$10^3/\mu L$
CAD	CAD = 1; Non-CAD = 0	Binary

CAD and non-CAD labels were taken from hospital diagnosis records. CAD cases were identified through assessments by consultant cardiologists, based on clinical examinations and additional investigations when needed. The specific tests or thresholds applied in individual cases and the criteria for non-CAD classification were not specified. Repeated visits were included as separate records when the CBC findings had changed. Therefore, the 3,301 records may represent fewer than 3,301 unique patients.

Data Preparation and Exploratory Analysis:

Before model development, descriptive analysis was carried out to examine the distribution of the dataset variables. The distributions of records by gender and age were examined, along with those of the numerical features. Correlation analysis was also applied to examine the relationships among the predictor variables and CAD status.

Numerical features were standardized using StandardScaler from scikit-learn. This method subtracts the mean of each feature and divides by its standard deviation. The scaling procedure was not documented in sufficient detail to establish whether the scaler was fitted separately within the training data and each cross-validation fold. Therefore, the possibility of preprocessing-related information leakage remains a limitation.

Machine Learning Models and Validation:

The dataset was divided into training and testing sets using an 80:20 split. From the entire dataset, 2,640 records were used for model training and 661 records formed the test set. Five-fold cross-validation was used within the training data during model development, and the held-out test set was reserved for final performance evaluation. Six classifiers were evaluated: Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB). The 80:20 split was performed at the record level, without grouping multiple visits from the same patient. The dataset was split at the record level rather than the patient level.

Data cleaning and preprocessing were carried out in Python using pandas within a Jupyter Notebook environment. Feature scaling and the six machine-learning classifiers were implemented using scikit-learn. Exact Python and library versions, classifier hyperparameters, random seed, and tuning settings were not recorded.

Performance Measures:

Model performance was evaluated using accuracy, precision, recall/sensitivity, specificity, and F1-score. Accuracy represents the overall proportion of correctly classified records. Precision measures the proportion of records predicted as CAD that were actually labelled as CAD, while recall (sensitivity) measures the proportion of CAD records correctly

identified by the model. Specificity measures the proportion of non-CAD records correctly identified and was calculated from the confusion-matrix counts as TN/(TN + FP). The F1-score provides a combined measure of precision and recall.

Confusion matrices were also used to summarize true positive, true negative, false positive, and false negative predictions for each classifier. All final model comparisons were based on performance on the held-out test data.

Ethical Considerations and Data Privacy:

The study was retrospective and involved no direct interaction with patients. Direct patient-identifying information was removed before analysis. Formal documentation of ethics approval, exemption, or an informed-consent waiver determination is not available for this study; therefore, no approval, exemption, or waiver is claimed.

Results:

Dataset Characteristics:

The final dataset contained 3,301 records, including 1,607 CAD-labelled records and 1,694 non-CAD records. Among the included records, 1,938 were coded as female and 1,363 as male. Of the female-coded records, 841 were labelled as CAD and 1,097 as non-CAD. Among male-coded records, 766 were labelled as CAD and 597 as non-CAD.

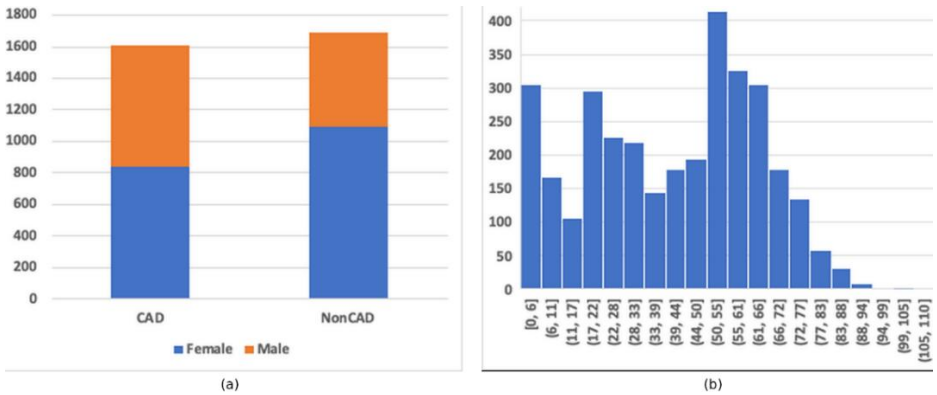


Figure 2. (a) Gender-wise breakup of the dataset (b) Age distribution of the study population.

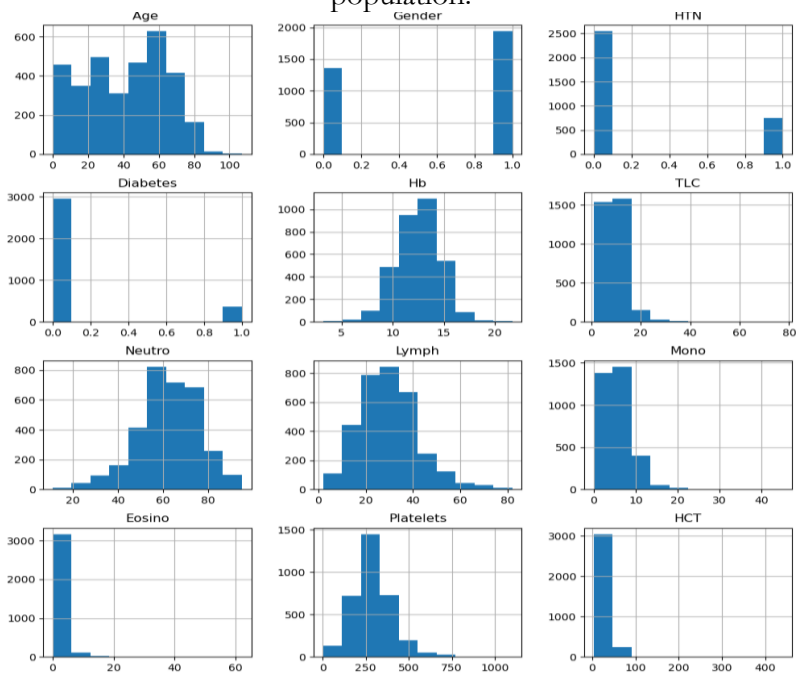


Figure 3. Distribution of the predictor variables

The age distribution of the records is shown in Figure 2(b). The dataset included records from a wide age range. The largest number of records was observed in the 50–55-year age group, while a large proportion of records were between approximately 50 and 70 years of age.

The distributions of the remaining predictor variables are presented in Figure 3. The variables showed different ranges and distribution patterns, which also supported the use of feature scaling before model development.

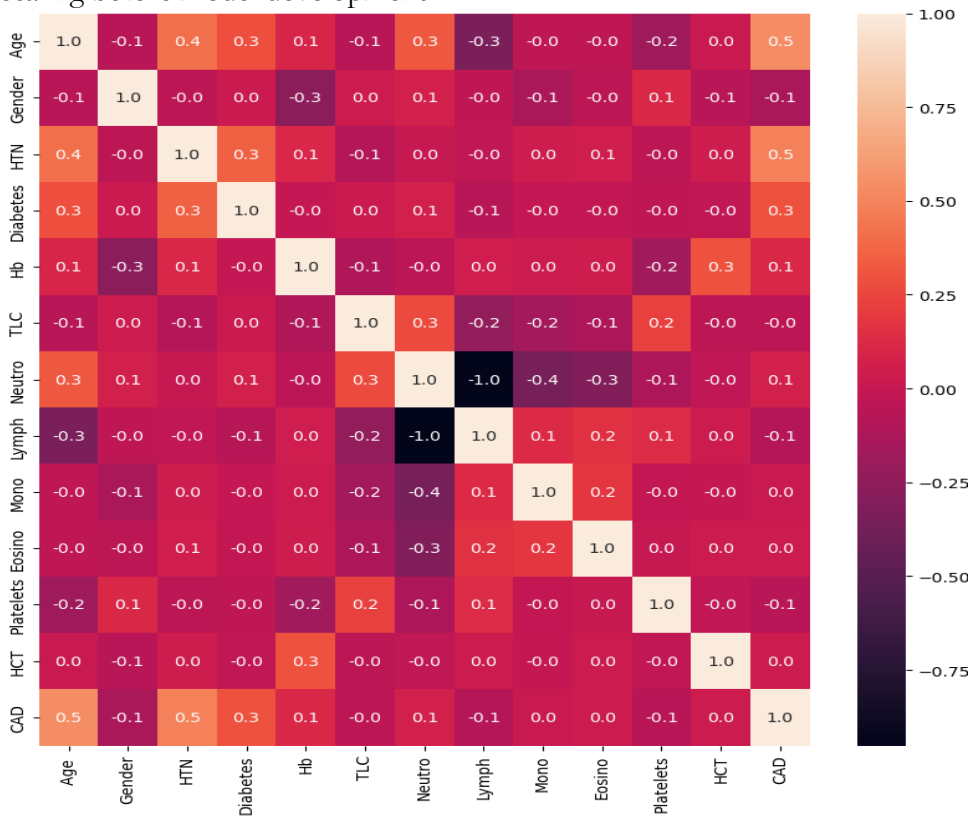


Figure 4. Correlation matrix of the predictor variables and CAD status

Machine Learning Performance:

Evaluation of all six machine learning classifiers used the 661 records in the held-out test set. Each classifier's true-positive, true-negative, false-positive, and false-negative counts were used to calculate its performance measures.

Table 2. Performance of the machine learning classifiers on the held-out test set

Model	Accuracy (%)	Precision (%)	Recall/ Sensitivity (%)	Specificity (%)	F1-score (%)
Logistic Regression	80.33	**84.72**	75.22	85.71	79.69
K-Nearest Neighbours	78.67	81.73	75.22	82.30	78.34
Random Forest	**81.24**	84.13	78.17	84.47	81.04
Decision Tree	75.64	77.47	74.04	77.33	75.72
Support Vector Machine	80.94	84.03	77.58	84.47	80.67
Naïve Bayes	80.79	82.32	**79.65**	81.99	80.96

Figure 5 compares accuracy, precision, recall, and F1-score across the six models. Logistic Regression, Random Forest, Support Vector Machine, and Naïve Bayes showed relatively close performance. Decision Tree had lower values on most of these metrics.

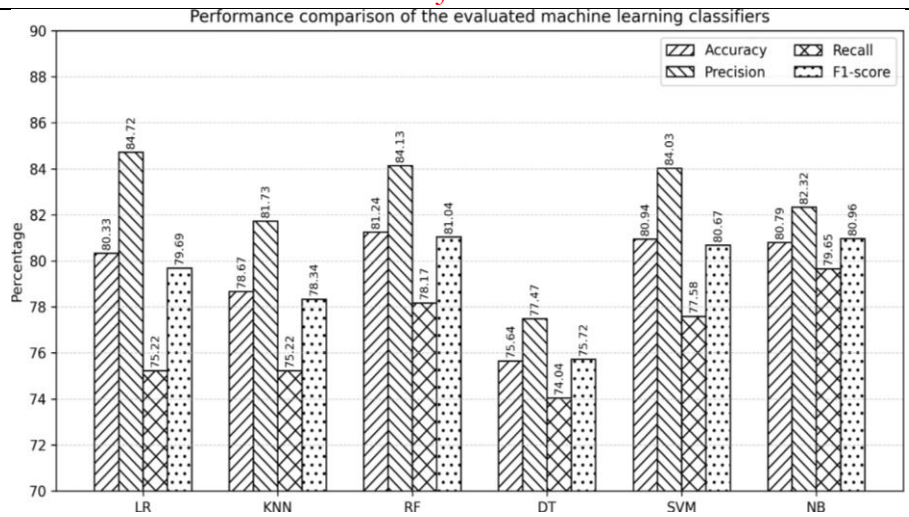


Figure 5. Performance comparison of the evaluated machine learning classifiers.

As shown in Figure 5, Random Forest achieved the highest reported overall accuracy and F1-score, while Logistic Regression produced the highest precision. Naïve Bayes achieved the highest recall/sensitivity. Several estimates were numerically close; without uncertainty estimates, these differences should be interpreted descriptively rather than as evidence of statistical superiority.

Confusion Matrix Analysis:

The confusion matrices provide further information about the classification errors made by the individual models. Logistic Regression correctly classified 276 non-CAD and 255 CAD records, while 46 non-CAD records were classified as CAD and 84 CAD records were classified as non-CAD. The KNN model correctly identified the same number of CAD cases but produced a larger number of false-positive predictions.

Random Forest correctly classified 272 non-CAD and 265 CAD records, with 50 false positives and 74 false negatives. Support Vector Machine produced a similar pattern, with 272 true negatives and 263 true positives. Naïve Bayes correctly identified the largest number of CAD cases., with 270 true positives and 69 false negatives, which is consistent with its higher recall.

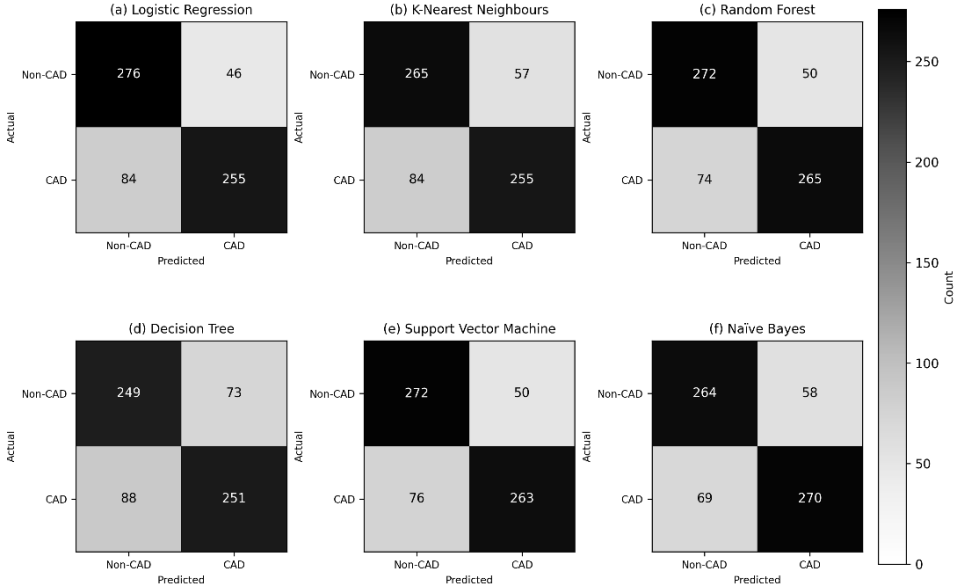


Figure 6. Confusion matrices for the six machine learning classifiers: (a) Logistic Regression, (b) K-Nearest Neighbours, (c) Random Forest, (d) Decision Tree, (e) Support Vector Machine, and (f) Naïve Bayes.

The confusion matrices therefore show some difference in the type of errors made by the classifiers. Logistic Regression produced higher precision, while Naïve Bayes identified a greater proportion of the CAD cases. Random Forest provided the strongest overall balance between the evaluated performance measures.

Figure 6 presents the confusion matrices for all six classifiers on the held-out test set. The matrices show differences in the balance of true-positive, true-negative, false-positive, and false-negative predictions across the models.

The confusion matrices show that the classifiers differed mainly in the type of classification errors they produced. Logistic Regression produced fewer false-positive predictions and therefore achieved the highest precision. Naïve Bayes correctly identified the largest number of CAD cases and produced the highest recall. Random Forest provided the strongest overall balance between correct CAD and non-CAD classifications, which is also reflected in its higher accuracy and F1-score.

Discussion:

Main Findings:

This study examined whether a restricted set of routinely collected CBC parameters together with age, gender, hypertension, and diabetes could provide useful information for CAD classification. The results show that several conventional machine learning models achieved accuracy values close to 80%. Random Forest produced the highest accuracy and F1-score, although Support Vector Machine, Naïve Bayes, and Logistic Regression showed only small differences in overall performance.

The main finding is therefore not that one machine learning algorithm was clearly superior to the others. Rather, the models retained useful discriminatory ability even when their inputs were limited to this small combined feature set. Several classifiers produced numerically similar estimates; without uncertainty analysis, this does not establish model equivalence. Because separate basic-clinical-only and CBC-only models were not evaluated, the independent or added contribution of CBC cannot be determined from this study.

False-negative classifications are potentially more concerning because CAD-labelled records are classified as non-CAD, whereas false-positive classifications may lead to unnecessary further assessment. CBC measurements are also nonspecific and can be affected by infection, inflammation, anemia, medications, and other medical conditions. The evaluated classifiers should therefore be regarded as exploratory and not as replacements for established CAD diagnostic procedures.

Comparison with Previous Studies:

Many previous studies have reported higher predictive performance for CAD and heart disease than that observed in the present study [15][19][17][26][27]. However, direct comparison of accuracy values is difficult because these studies used different patient populations, feature sets, outcome definitions, and validation procedures. Some studies also used ECG, echocardiography, imaging findings, symptoms, and a wider range of laboratory or clinical variables.

The present study intentionally used a more restricted set of predictors. This difference is important when interpreting the approximately 80–81% classification accuracy obtained by the better-performing models. The objective was not to maximize classification performance using all available clinical information, but to examine the performance attainable under this restricted information set.

Recent research gives further support to this approach. Koloi et al. [10] investigated early-stage CAD using routine clinical biomarkers, while [11] used hematological markers for acute myocardial infarction classification. [12] also included routine hematological variables such as white blood cell count, hemoglobin, hematocrit, neutrophils, and platelets in models for premature CAD.

Other studies have specifically demonstrated the usefulness of routine blood information in cardiovascular prediction. Hematological predictors were examined for cardiovascular risk by [8], and a CBC-derived score for acute myocardial infarction was developed by [9]. [13] more recently reported that adding routine blood-test variables to baseline clinical information improved cardiovascular classification. The outcomes differ from those in the present study, preventing direct comparison of results. These studies nonetheless support the view that routine blood measurements may contain useful cardiovascular information.

Implications and Future Research:

These findings support further investigation of routinely available CBC and basic clinical variables as supplementary inputs for CAD-related assessment, rather than as stand-alone diagnostic tools. Independent and multicentre datasets should be used in future studies to compare basic-clinical-only, CBC-only, and combined models and to evaluate uncertainty and calibration. Prospective validation should also precede consideration of clinical application.

Limitations:

Several limitations should be considered when interpreting these findings. First, the data were collected from a single hospital, and the characteristics of patients attending this hospital may differ from those in other settings. External validation is therefore needed. Second, complete-case filtering reduced approximately 22,400 candidate records to 3,301 records and may have introduced selection bias. Third, the study did not compare basic-clinical-only, CBC-only, and combined feature sets. Therefore, the independent or incremental contribution of CBC parameters cannot be determined. Finally, external validation and model calibration were not performed. The findings should therefore be considered exploratory. Specific diagnostic tests and criteria, detailed exclusion counts, outlier thresholds, exact model settings, and software versions were not available, which limits reproducibility. Records were split without grouping repeated visits by patient, so the same patient may have contributed observations to both subsets and the reported performance may be optimistic. Since repeated visits may not be independent, confidence intervals calculated directly from aggregate confusion-matrix counts would require assumptions that could not be verified. ROC-AUC values were not calculated because the continuous model scores were unavailable. External validation and model calibration are required before the findings can be considered for clinical use.

Conclusion:

This retrospective study explored the use of complete blood count parameters and a small number of basic clinical variables for machine learning-based classification of coronary artery disease. The analysis included 3,301 hospital records and compared six commonly used classifiers.

Random Forest achieved the highest overall accuracy of 81.24% and F1-score of 81.04%, while Logistic Regression produced the highest precision and Naïve Bayes produced the highest recall/sensitivity. Several models differed only slightly in performance, suggesting that the restricted combined feature set allowed useful retrospective discrimination. These findings establish neither that CBC alone predicts CAD nor that it adds value beyond the basic clinical factors.

The results remain exploratory and should not be taken as evidence of a clinically validated diagnostic system. Future studies should evaluate this approach using independent and multicentre datasets with clearly defined clinical outcomes, stronger validation procedures, and calibration. Comparing basic-clinical-only, CBC-only, and combined models should also allow the added contribution of hematological variables to be quantified.

References:

[1] "Cardiovascular diseases (CVDs)." Accessed: Nov. 10, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE J. Res.*, vol. 68, no. 4, pp. 2488–2507, 2022, doi: 10.1080/03772063.2020.1713916;.

[3] R. Alizadehsani *et al.*, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Comput. Biol. Med.*, vol. 111, p. 103346, Aug. 2019, doi: 10.1016/J.COMPBIOMED.2019.103346.

[4] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association," *Circulation*, vol. 147, no. 8, pp. E93–E621, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[5] "About Coronary Artery Disease (CAD) | Heart Disease | CDC." Accessed: Nov. 10, 2026. [Online]. Available: https://www.cdc.gov/heart-disease/about/coronary-artery-disease.html

[6] "(PDF) Z-Alizadeh Sani Dataset." Accessed: Nov. 10, 2026. [Online]. Available: https://www.researchgate.net/publication/296485502_Z-Alizadeh_Sani_Dataset

[7] "Heart disease - Symptoms and causes - Mayo Clinic." Accessed: Nov. 10, 2026. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[8] J. G. Truslow *et al.*, "Cardiovascular Risk Assessment Using Artificial Intelligence-Enabled Event Adjudication and Hematologic Predictors," *Circ. Cardiovasc. Qual. Outcomes*, vol. 15, no. 6, p. E008007, Jun. 2022, doi: 10.1161/CIRCOUTCOMES.121.008007.

[9] H. Zhang, Z. Luo, D. Jia, D. Li, Y. Jia, and Z. Wan, "A risk score derived from complete blood count contributes to early risk stratification of acute myocardial infarction at the emergency department," *Int. Emerg. Nurs.*, vol. 68, May 2023, doi: 10.1016/j.ienj.2023.101287.

[10] A. Koloi *et al.*, "Predicting early-stage coronary artery disease using machine learning and routine clinical biomarkers improved by augmented virtual data," *Eur. Hear. J. - Digit. Heal.*, vol. 5, no. 5, pp. 542–550, Sep. 2024, doi: 10.1093/EHJDH/ZTAE049.

[11] "(PDF) Predicting acute myocardial infarction from haematological markers utilizing machine learning and explainable artificial intelligence." Accessed: Nov. 09, 2026. [Online]. Available: https://www.researchgate.net/publication/379271023_Predicting_acute_myocardial_infarction_from_haematological_markers_utilizing_machine_learning_and_explainable_artificial_intelligence

[12] Y. C. He *et al.*, "Construction of diagnostic and prognostic models for premature coronary artery disease based on multiple machine algorithms," *Front. Cardiovasc. Med.*, vol. 12, 2025, doi: 10.3389/FCVM.2025.1611709.

[13] N. Tasmurzayev *et al.*, "Enhancing Cardiovascular Disease Classification with Routine Blood Tests Using an Explainable AI Approach," *Algorithms 2025, Vol. 18, Page 708*, vol. 18, no. 11, p. 708, Nov. 2025, doi: 10.3390/A18110708.

[14] 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," *Eur. Hear. journal. Digit. Heal.*, vol. 6, no. 1, pp. 7–22, Jan. 2024, doi: 10.1093/EHJDH/ZTAE080.

[15] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, Oct. 2019, doi:

10.1016/J.CMPB.2019.104992.

[16] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. Benjula Anbu, P. Jayagopal, and G. T. Dalu, "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database," *Sci. Reports 2023 131*, vol. 13, no. 1, pp. 13588-, Aug. 2023, doi: 10.1038/s41598-023-40717-1.

[17] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci. 2020 16*, vol. 1, no. 6, pp. 345-, Oct. 2020, doi: 10.1007/S42979-020-00365-Y.

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

[19] R. Alizadehsani *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004.

[20] C. A. ul Hassan *et al.*, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," *Sensors (Basel).*, vol. 22, no. 19, p. 7227, Oct. 2022, doi: 10.3390/S22197227.

[21] H. Khdair and N. M. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 28–36, May 2021, doi: 10.14569/IJACSA.2021.0120505.

[22] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," *2022 Adv. Sci. Eng. Technol. Int. Conf. ASET 2022*, 2022, doi: 10.1109/ASET53988.2022.9734880.

[23] K. N. Devi, S. Suruthi, and S. Shanthi, "Coronary Artery Disease prediction using Machine Learning Techniques," *8th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2022*, pp. 1029–1034, 2022, doi: 10.1109/ICACCS54159.2022.9785140.

[24] "Framingham Heart Study." Accessed: Nov. 10, 2026. [Online]. Available: https://www.framinghamheartstudy.org/

[25] Q. Li, A. Campan, A. Ren, and W. E. Eid, "Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system," *Int. J. Med. Inform.*, vol. 163, Jul. 2022, doi: 10.1016/j.ijmedinf.2022.104786.

[26] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine Learning Predictive Models for Coronary Artery Disease," *SN Comput. Sci.*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/S42979-021-00731-4.

[27] C. Wang *et al.*, "Development and Validation of a Predictive Model for Coronary Artery Disease Using Machine Learning," *Front. Cardiovasc. Med.*, vol. 8, Feb. 2021, doi: 10.3389/FCVM.2021.614204.

[28] S. M. Naushad, T. Hussain, B. Indumathi, K. Samreen, S. A. Alrokayan, and V. K. Kutala, "Machine learning algorithm-based risk prediction model of coronary artery disease," *Mol. Biol. Rep.*, vol. 45, no. 5, pp. 901–910, Oct. 2018, doi: 10.1007/S11033-018-4236-2.

[29] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," *Health Soc. Care Community*, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.