STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, October 2025, pp 1656–1675.
Published online in International Academic Press (www.IAPress.org)

Analyzing and Classifying Coronary Artery Disease Severity Using Statistical Methods and Machine Learning Techniques

Meriem Bounekdja [1], Soumia Kharfouchi [2,*], Abdennour Boulesnane [2]

[1] *Department of Mathematics, Faculty of Exact Sciences, Frères Mentouri Constantine University, Constantine, Algeria*
[2] *BIOSTIM Laboratory, Faculty of Medicine, University Salah Boubnider Constantine 3 ,Algeria*

Abstract **Background:** Metabolic Syndrome (MS) is a cluster of risk factors, including large waist size (LWS), high blood pressure (HBP), high cholesterol levels (HDL), high blood glucose (HBG), glycemic index (GI), and hypertriglyceridemia (HTG), which collectively increase the risk of developing cardiovascular diseases such as Coronary Artery Disease (CAD). Understanding the relationship between MS and CAD severity is crucial for developing targeted prevention and treatment strategies.
Methods: This study conducted an etiological and descriptive analysis to characterize the profiles of CAD patients with MS using various statistical methods. These methods included correlation analysis and odds ratio calculations to evaluate the significance of MS components. Multiple machine learning (ML) models, including Multilayer Perceptron (MLP), Decision Tree (DT), Logistic Regression (LR), AdaBoost (ABT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and an ensemble Voting Classifier (VC), eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) were employed to classify CAD severity and identify modifiable risk factors specific to various MS combinations. The effectiveness of these models was evaluated and compared.
Results: The analysis identified HDL, HBG, and LWS as significant aggravating factors for CAD, while HTG appeared to be protective. The XGBoost model demonstrated superior predictive accuracy, achieving an accuracy of 83.12% in predicting CAD severity, compared to other ML models. The inclusion of MS features significantly enhanced the performance of all ML models.
Conclusions: The findings underscore the importance of incorporating comprehensive clinical features in predictive models for CAD. The study suggests that targeted prevention strategies and personalized treatment plans should consider the specific MS components influencing CAD severity. Future research should focus on validating these findings in larger, diverse populations and further integrating additional clinical and genetic data to refine predictive models.

Keywords Coronary Artery Disease, Coronary angiography, Metabolic Syndrome, Machine Learning, Ensemble Voting Classifier, Risk Factors.

AMS 2010 subject classifications 92C50, 68T05, 62H30

DOI: 10.19139/soic-2310-5070-2241

1. Introduction

Ischemic heart disease is a cardiac pathology. It is the most common cardiovascular disease and is one of the leading causes of death worldwide [38].

Coronary artery disease (CAD), also called ischemic heart disease or myocardial ischemia, corresponds to a decrease in the blood flow irrigating the heart muscle, leading to the appearance of coronary lesions in the walls of the vessels that supply the heart [30]. These lesions cause atherosclerosis, the blockage of the coronary arteries

*Correspondence to: Soumia Kharfouchi (Email: s_kharfouchi@yahoo.fr). BIOSTIM Laboratory, Faculty of Medicine, University Salah Boubnider Constantine 3 ,Algeria.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

that carry blood containing oxygen to the heart. This prevents the heart muscle from functioning properly, thus increasing the risk of cardiac arrest. Great scientific advances have been made in the last two decades in the field of the study of this coronary heart disease because of its seriousness for human health, in particular in terms of experimental studies and clinical trials which are based on mathematical models explaining the evolution of the disease [6], [7]. This being so, in many cases, ischemic heart disease can be avoided (or prevented) if:

- One acts on its modifiable risk factors, namely, LWS, HBP, HDL, HBG, GI, HTG.
- One minimizes the influence of its non-modifiable risk factor, namely, age and sex, by predicting and simulating appropriate variables that can describe the course of the disease according to these two factors.

This has led to prevention and regression trials. Prevention trials consist of acting on the risk factors to reduce the disease's incidence. For instance, coronary heart mortality has declined in the US [42] because, in addition to advances in the treatment of coronary heart disease, treatment of hypertension has been improved, tobacco consumption has been reduced [23], and diet changes have been implemented [33]. Regression trials, for their part, act on risk factors when the disease already exists.

In this framework, the combination of LWS and at least two modifiable risk factors is defined as MS. This poses a significant public health problem. It has been found that a certain proportion of patients with CAD present an MS. The individual components of MS known for predicting CAD angiographic severity are (LWS, HBP, HBG, GI, HDL, and HTG). They can be supplemented, reduced, or revised to have a much more subtle vision of the concept of risk. Hence, prevention can be more targeted (by acting on the most influential factors or constituting the most dangerous combinations). This begs the question: Is MS an aggravating factor concerning coronary lesions?
To answer this question, some studies have been done on the influence of MS on CAD. They mainly focused on the influence of each factor, taken alone, on the disease: let us quote those of [19] on the risk of CAD linked to diabetes, those of [11] on the influence of HBP on the CAD and those of [15] on cardiometabolic complications and their relation to obesity. Note that some similar studies also combined several risk factors, particularly those of [43] on evaluating the frequency of MS by describing its clinical and para-clinical aspects.

The influence of the MS on CAD is a topic that has been addressed by several statistical methods based on mathematical models that describe its behavior and that allow the interpretation of the studied phenomenon. Some studies relied mainly on descriptive statistics and standard (traditional) statistical methods to analyze the data by calculating the mean and the median for the quantitative variables and the numbers and the frequencies for the qualitative variables [43]. Moreover, a logistic regression was used for a case-control study of the prevalence. Using the SCOR model is interesting and useful for predicting the risk of death at any age and calculating the risk of fatal cardiovascular death [20]. Consistency tests helped to raise the question of queries by freezing the database after correcting all the queries [33].

However, despite the increasing complexity of data derived from CAD studies, most existing research remains relatively elementary. This is particularly noteworthy given the potential of computationally sophisticated models to learn complex, nonlinear functions essential for accurate data analysis. Supervised Machine Learning (ML) offers a promising alternative, capable of precisely predicting key parameters that directly and significantly influence the occurrence and severity of CAD. By leveraging these ML algorithms, we can analyze patient profiles at higher risk of severe CAD associated with MS, identifying patterns that are otherwise difficult to discern.

In the literature, supervised ML methods have demonstrated adaptability and effectiveness in various fields of medical monitoring and diagnosis, such as the prediction of infarcts [8], [2], the diagnosis of pulmonary pathologies [32], and cancers [27]. Specifically, in the context of CAD, several studies have highlighted ML algorithms' potential in diagnosing CAD and estimating associated risk factors. Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) have been particularly effective in these applications. For instance, Zhang et al. [46] employed ML algorithms with feature selection techniques to predict CAD, achieving high accuracy and efficiency. Similarly, Kolukisa et al. [26] improved diagnostic performance using ensemble feature selection and classification methods for CAD diagnosis. Yu et al. [45] utilized ML algorithms to detect CAD in its early stages, which is crucial for effective treatment and management. Moreover, in addition to diagnostic applications, ML algorithms have been employed to estimate risk factors related to heart diseases. Revathi et al. [36] used a Multilayer Perceptron Model to estimate heart disease risk factors, providing valuable insights for disease prevention and management.

Recent work by Eyupoglu et al. [14] shows that deep learning models (CNN and RNN-LSTM) outperform AUROC scores (0.89-0.93) compared to traditional logistic regression by capturing nonlinear relationships in cytokine biomarkers [37]. Also, Yu et al. [14] emphasizes XGBoost's enhanced predictive power for CAD risk stratification of 684 features through SHAP analysis, revealing complex interactions between HDL-C, blood pressure, and inflammatory markers that conventional models often oversimplify [21], [1]. These studies collectively support the claim that emerging techniques overcome traditional limitations in handling high-dimensional data and biological complexity. The following reasons partly explain the ML relatively success:

- They do not require any assumptions on the variables, unlike classical statistical methods.
- They are quite suitable for dealing with complex problems on which it is impossible a priori to specify the form of the relations between the variables used.
- These systems learn the relationships between variables from a data set, just like the human brain would. Thus, the network is self-configuring, based on the examples provided, which gives it adaptive power to deal with any change in situation, such as a database update.

Consequently, the primary objective of this paper is to comprehensively examine the relationship between MS and its components with the severity of CAD as assessed through angiographic measures. To achieve this, we conduct two epidemiological studies. The first study has an etiological and descriptive focus, aiming to better characterize the profiles of CAD patients with MS by analyzing various combinations of MS components. The second study involves the application of supervised ML prediction methods to identify modifiable CAD risk factors specific to each MS combination. Specifically, this study seeks to develop predictive models for the severity of CAD using logistic regression (LR), AdaBoost (ABT), K-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), Gaussian naive Bayes (GNB), multilayer perceptron (MLP), ensemble Voting Classifier (VC) techniques, eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM).
The output variable in these models measures the degree of severity of lesions as observed through coronary angiography, with severity levels ranging from 1 to 7. This measurement allows for the analysis of the influence of various MS-related factors on CAD severity. This comprehensive study aims to identify the most effective predictive model among those proposed to accurately determine CAD angiographic severity in previously unseen patients (i.e., those not included in our training set). Ultimately, the proposed model will provide diagnostic predictions of CAD severity levels based on the specific characteristics of individual patients.

2. Materials and methods

2.1. Metabolic syndrome

The concept of metabolic syndrome (MS), initially named X syndrome, was issued in 1988 by Gerald Reaven to designate a pivotal abnormality that is both diabetogenic and atherogenic [35]. It is also known as insulin resistance syndrome, which directly contributes to cardiovascular and/or CAD risk. Indeed, the Interheart Study is a case-control study carried out in 52 developing and developed countries, showing that the five factors, smoking, HBP, hypercholesterolemia, and abdominal obesity alone, explain 80% of cases of first infarction. In particular, obesity, characterized by an abdominal visceral distribution of excess body fat, is an inseparable factor of MS. Thereby, the National Cholesterol Education Program (NCEP) and the Third Adult Treatment Panel (ATP III) modified the waist criterion into a body mass index (BMI) of more than 25 kg/m^2. However in our study, to be diagnosed with MS one should present an LWS (for men 40 inches or more, for women 35 inches or more) plus two, at least, of the following five symptoms: HBP (130/85mmHg or greater); HBG (100mg/dL or more); GI (a step that precedes type 2 diabetes); HDL ($<$ 40 mg/dL in men and $<$ 50 mg/dL in women); HTG (150mg/dL or higher).

2.2. CAD angiographic severity

CAD is diagnosed when luminal diameter stenosis of at least one major epicardial coronary artery, is higher than 50% [22]. There are several ways to calculate CAD severity. The present study used coronagraphy as the only criterion to judge the severity of CAD angiographic. Stress tests and scintigraphy were not considered because, on

the one hand, they are less revealing than coronary angiography; on the other hand, many patients have not done them.

Coronary angiography measures the degree of severity of the lesions. It gives results that graduated from 1 to 7, in order of increasing severity (1 means there are no lesions, and 7 is the most severe lesion). The severity score was determined for each patient according to the opinion of the experienced cardiologists of the cardiology department of the Constantine Military Hospital. It is against this variable that we measure the influence of the various factors. In our study, patients with angiographic severity scores of 1 or 2 are classified as Non-Aggravating, while those with scores from 3 to 7 are considered Aggravating.

2.3. Data description

We conducted a study in the field of Cardiac Surgery, collecting clinical data from 1670 patients (1295 men, 61.08 $\pm$ 9.38 years; 375 women, 61.04 $\pm$ 9.37 years) at the Cardiology Department of the Military Hospital in Constantine. All patients were diagnosed with CAD, and 820 of them (49.10%; 95% confidence interval: [46.7-51.5]) presented with MS, a well-known clinical marker of increased CAD risk. Given the diverse and cumulative nature of MS components, our study aimed to investigate the risk associated with different combinations of these components to evaluate CAD severity better. As shown in Table 1, the factors analyzed in our study included sex, age, ischemia, smoking, BMI, clinical condition, ECG status, heredity, LWS, HBP, HDL, HBG, HTG, GI, and CAD angiographic severity (the target variable).

Table 1. Description of dataset variables.

Variables		**Descriptions**
Age		Patients Age (in Years)
Sex		0 : male and 1 : female
Smoking		0 : no and 1 : yes
BMI		Body Mass Index (kg/m^2)
Heredity		0 : no and 1 : yes
Clinical		0 : Angor, 1 : SCA+, 2 : SCA-, 3 : SCA-tr+, 4 : IMS, and 5 : IVG/IDM
Normal ECG		0 : no and 1 : yes
Metabolic Syndrome Combinations	Ischemia	0 : no and 1 : yes
	LWS	0 : no and 1 : yes
	HBP	0 : no and 1 : yes
	HBG	0 : no and 1 : yes
	GI	0 : no and 1 : yes
	HDL	0 : no and 1 : yes
	HTG	0 : no and 1 : yes
CAD angiographic severity		0 : Non-Aggravating and 1 : Aggravating

2.4. Statistical analysis

Data are presented as means $\pm$SD for continuous clinical variables, while for discrete or categorical variables, they are presented as frequencies. Prevalences and odds ratios (OR) (95% confidence interval (CI)) are estimated to see

the influence of different combinations of MS components on the severity of a CAD. An odds ratio greater than 1 means that the combination has an "aggravating" influence on the development of CAD. An odds ratio of less than 1 means that the combination has a "protective" influence on the appearance of a CAD. Comparison among groups was performed using Pearson's chi-square test for discrete or categorical variables. A correlation matrix is used to assess the dependence between several variables simultaneously. In our case, the Pearson correlations matrix is used for the dimension reduction of the variable's space and the elimination of the irrelevant variables since variables that carry, to a certain extent, the same information are strongly correlated.

However, we will not be able to express an opinion on certain variables that are considered undecided factors. Thus, further analysis is essential to remove the indecision and make an accurate and effective prediction. For this, we propose the use of supervised ML techniques. These techniques provide models whose complexity is adjustable according to the data through learning from them. They are also known to be good candidates for prediction tasks.

2.5. Feature selection and data balancing

Feature selection techniques employ criteria to assign scores to features, indicating their importance in determining target labels [18]. In disease diagnosis, this helps identify crucial features for accurate diagnoses, assisting physicians in choosing more effective treatment strategies [41].

Although the attributes count is already low in our collected data (only eight features), ensuring that only pertinent features are utilized while removing any irrelevant (or redundant) variables is always advantageous, enhancing training efficiency and mitigating overfitting.

This study applied a feature selection method to eight input features from a dataset comprising 1670 instances. To evaluate feature importance, five ML regressors from the scikit-learn library [28], including AdaBoostRegressor, BayesianRandomForestRegressor, GradientBoostingRegressor, DecisionTreeRegressor, ExtraTreesRegressor and SHAPvaluesRegressors were employed.

On the other hand, to enhance the performance of the proposed CAD classification model, we addressed the issue of data imbalance within our dataset, which consisted of 1670 instances comprising 1231 Aggravating and 439 Non-Aggravating cases.

For this purpose, we employed the Synthetic Minority Over-sampling Technique (SMOTE), that generates synthetic samples for the minority class by interpolating between existing instances. This technique preserves the underlying data distribution while avoiding duplication. Additionally, we implemented a Balanced Random Forest (BRF) model, an ensemble method designed to handle imbalanced data by balancing class weights during tree construction.

2.6. Conventional machine learning models

To classify CAD severity as either Aggravated or Non-aggravated, we investigate a range of conventional ML models. In our classification, Non-Aggravating refers to patients with CAD angiographic severity scores of 1 or 2, whereas Aggravating includes those ranging from 3 to 7. The primary objective is to assess the performance of these models and determine the most effective approaches for accurately categorizing CAD severity in medical datasets.

2.6.1. Multilayer perceptrons Multilayer perceptrons (MLP) are the most common and simplest neural models, often used for supervised learning tasks [3]. These models consist of several layers, with neurons receiving signals from various sources, typically input variables denoted as x in $\mathbb{R}^p$. Each input is weighted, with the weight representing its importance; higher weights correspond to stronger signals and more influential inputs.

The computation within a neuron is characterized by a transfer function $y = g(A)$, where A is a function mapping from $\mathbb{R}^p$ to $\mathbb{R}$ and g is an activation function mapping from $\mathbb{R}$ to $\mathbb{R}$. Typical activation functions comprise the sigmoid, hyperbolic tangent, and rectified linear unit (ReLU) functions. These functions are crucial for introducing non-linearities necessary to capture intricate data relationships during learning.

2.6.2. Decision Tree Decision Tree (DT) is a well-established supervised learning technique in ML. They begin with a root node representing the entire dataset, branching into multiple nodes based on chosen attributes that best

divide the data [39]. Decisions are made based on specific features at each internal node, recursively partitioning the data until a stopping criterion is met. Leaf nodes hold outcomes: class labels for classification or numerical values for regression.
The tree's structure allows for straightforward interpretation, with each path representing a sequence of decisions leading to an outcome. Building a decision tree involves starting with a main decision, adding chance nodes, and analyzing outcomes' expected values and probabilities.

2.6.3. Logistic Regression Logistic regression (LR) is a statistical technique primarily employed for binary classification tasks [31]. LR method estimates the probability of a specific input x belonging to a particular class through a logistic function. This is achieved through the logistic function given by:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1}$$

The coefficients β_0 and β_1 are determined from training data, allowing LR to model the association between the input features and the binary outcome. This model encapsulates information within its coefficients, which scale with the dimensionality, irrespective of the available number of observations. This characteristic makes it indicative of parametric methodologies.

2.6.4. Support Vector Machine SVM is a robust supervised learning algorithm extensively employed for classification and regression tasks in various fields, including medical diagnosis. SVM operates by identifying the optimal hyperplane that effectively separates different classes within the dataset while maximizing the margin between them [40]. This hyperplane serves as the decision boundary, enabling us to classify new instances based on their features accurately. SVM utilizes the following equation to find the optimal hyperplane for classification:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right) \tag{2}$$

Where $f(x)$ is the decision function that predicts the class label of a new instance x is determined by utilizing the support vectors and their corresponding Lagrange multipliers obtained during training. Here, N represents the number of support vectors, x_i represents each support vector, y_i denotes the class labels of the support vectors, α_i are the Lagrange multipliers obtained during training, and $K(x, x_i)$ is the kernel function, which calculates the similarity between x and the support vectors x_i, and b is the bias term. This equation encapsulates the essence of SVM by computing the decision function based on the support vectors and their corresponding Lagrange multipliers, adjusted by the kernel function and bias term.

2.6.5. Gaussian Naive Bayes Gaussian Naive Bayes (GNB) is a probabilistic classifier based on Bayes' theorem with the "naive" assumption that features are independent given the class label. The GNB classifier involves calculating the posterior probability of a class given a set of features using Bayes' theorem, assuming the likelihood of the features, given the class, a Gaussian distribution [9].
The equation used to perform the Gaussian Naive Bayes classification task is represented as:

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, \ldots, x_n)} \tag{3}$$

Where $P(y \mid x_1, x_2, \ldots, x_n)$ represents the posterior probability of class y given the features $x_1, x_2, \ldots, x_n$, $P(y)$ is the prior probability of class y, $P(x_i \mid y)$ is the likelihood of feature xi given class y, and $P(x_1, x_2, \ldots, x_n)$ is the marginal probability of the features.

2.6.6. K-Nearest Neighbors KNN is a non-parametric method renowned for its simplicity and intuitive nature in classification tasks. Since KNN is a non-parametric technique, no assumptions are made about the underlying data distribution, allowing it to adapt flexibly to diverse datasets without imposing strict structural constraints [47].

Furthermore, KNN identifies its nearest neighbors through a chosen distance metric, typically Euclidean distance, in the feature space. By leveraging the class labels of these neighbors, KNN assigns the most prevalent class label among them to the new data point.

2.6.7. Adaboost model Adaboost, abbreviated from Adaptive Boosting, is a favored ensemble learning technique used in classification and regression tasks. It combines multiple weak learners, typically decision trees, to create a strong learner [10]. Let T denote the number of weak learners, α_t represent the weight assigned to the t-th weak learner, and $h_t(x)$ be the t-th weak learner's prediction for input x. The final prediction $F(x)$ made by the Adaboost model for input x is given by:

$$F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{4}$$

Where sign(.) is the sign function, which outputs +1 for positive values and -1 for negative values. During the training process, Adaboost adjusts the weights α_t and updates the distribution of training samples to prioritize misclassified samples. After training, the model makes predictions based on the weighted combination of the weak learners' predictions.

2.6.8. XGBoost model XGBoost is a highly scalable and efficient gradient boosting implementation for performance and speed in supervised learning tasks. It proposes an ensemble of weak learners—typically decision trees—by sequentially optimizing a loss function using gradient descent and additive modeling [29]. At each iteration, a new tree is fitted to the residual error of the loss function, effectively correcting the errors made by previous trees.

Regularization techniques (L1 and L2) to prevent overfitting, tree pruning strategies to enhance generalization, and a sparsity-aware algorithm for handling missing data are key innovations of XGBoost. Additionally, parallel processing and cache optimization contribute to its fast training speed.

2.6.9. LightGBM model LightGBM is a gradient boosting framework for large datasets with high-dimensional features [30]. In this technique, A set of decision trees trained sequentially to minimize a loss function is proposed. However, LightGBM has two key innovations: Gradient-based One-Side Sampling and Exclusive Feature Bundling. GOSS accelerates training by retaining instances with large gradients while randomly sampling from the rest, whereas EFB reduces dimensionality by bundling mutually exclusive features.

Another distinctive feature of LightGBM is that it selects the leaf with the highest loss to split, leading to deeper trees that better capture complex patterns. This often enhances performance on structured datasets, While this can lead to overfitting if not properly regularized.

2.7. Proposed ensemble learning model

Ensemble learning is a powerful technique that combines multiple classifiers to improve prediction accuracy and robustness over individual models [13]. By aggregating the strengths of various ML models, ensemble methods can enhance performance, especially in complex tasks like CAD severity classification.

In our study, we propose an ensemble learning model incorporating three diverse ML classifiers based on their effectiveness demonstrated in our experimental analysis, which includes among others AdaBoost, K-Nearest Neighbors, and Decision Tree. Each classifier brings unique strengths, allowing the ensemble to capitalize on their complementary capabilities. To integrate these models, we utilize a Voting Classifier (VC) [29], which synthesizes the predictions from each base classifier to produce a final output. Specifically, we adopt a soft voting approach, where the predicted probabilities for each class are averaged across the base classifiers. The class with the highest average probability is then selected as the final prediction. This method not only enhances the robustness of our model but also reduces the likelihood of overfitting to individual classifiers, ultimately leading to improved accuracy in classifying CAD severity.

2.8. Model performance criterias

This study uses four key assessment metrics, accuracy, precision, recall, and F1-score, to evaluate the model's performance [17].
Each metric is measured on a scale from 0.0 (reflecting poor performance) to 1.0 (indicating optimal performance). These metrics are averaged across ten folds. A higher metric value indicates a higher level of model efficiency.
Accuracy denotes the fraction of correct predictions relative to all predictions made. This metric is formally characterized as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision is the quotient of true positive predictions divided by the sum of true positive and false positive predictions. Its definition is:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall quantifies the proportion of true positive predictions relative to the total of true positive and false negative predictions. Its calculation is outlined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Specificity evaluates the fraction of negative cases that were accurately classified, computed as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

The F1-score is expressed as the harmonic mean of precision and recall, with its definition as follows:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

In the equations mentioned earlier, TP, FP, TN, and FN signify True Positive, False Positive, True Negative, and False Negative counts, respectively. True Positives (TP) represent accurately identified Aggravating cases, True Negatives (TN) denote Non-Aggravating cases correctly identified as such, False Positives (FP) indicate Non-Aggravating cases mistakenly labeled as Aggravating, and False Negatives (FN) are Aggravating cases inaccurately identified as Non-Aggravating.

On the other hand, the correlation coefficient is used to measure the linear correlation between the actual or measured value and the calculated value. Its formula is:

$$R = \frac{\sum_i (X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2} \times \sqrt{\sum_i (Y_i - \bar{Y})^2}} \tag{10}$$

With X_i = measured value; Y_i = calculated value; $\bar{X}$ = mean of the measured values; $\bar{Y}$ = mean of the calculated values.
According to [25], by convention, the relation is perfect if $R = 1$; very strong if $R > 0.8$; strong if R is between 0.5 and 0.8; of medium intensity if R is between 0.2 and 0.5; weak if R is between 0 and 0.2; null if $R = 0$. In general, if R is less than 0.7, then the relation is subject to problem [4].

2.9. Gender Bias Mitigation and Stratified Evaluation

To assess potential gender disparities, we trained a single XGBoost model on the balanced dataset (SMOTE-augmented) and evaluated its performance separately on male and female subpopulations. Statistical parity was ensured by maintaining proportional representation of both genders during resampling. We quantified fairness using equalized odds (difference in F1 scores between genders) and report performance metrics stratified by sex.

2.10. Machine Learning Implementation and cross-validation

All ML models were implemented using custom Python code executed on Google Colab. Computations were run on i7 Dell laptop equipped with a 12th generation and 32GB RAM, running Ubuntu 20.04.

We selected hyperparameters through 10-fold stratified cross-validation as depicted in Table 2, to ensure robustness and to account for class imbalance. This strategy maintains the original class proportions in each fold. Particularly, the Stratified-K-Fold method, from the scikit-learn library, was applied, under n-splits=10, shuffle=True and a random-state=42 to ensure reproducibility. Classification accuracy was the primary evaluation metric, while secondary metrics such as precision, recall, F1-score, and the area under the ROC curve (AUC) were also reported.

3. Results

In this section, we present the findings from our analysis of the relationship between MS and CAD angiographic severity using statistical methods. Additionally, we detail the performance of various ML models in classifying CAD severity as either Aggravated or Non-aggravated. The results encompass descriptive statistics, correlation and regression analyses, and the efficacy of different ML techniques, including feature selection and data balancing. We also comprehensively compare and evaluate the ML models employed in this study.

3.1. Statistical analysis of the relationship between MS and CAD severity

First, we will see the possibility of the existence of a link between MS and the CAD angiographic severity in the two subpopulations of patients with MS and without MS. Figure 1 shows the distributions of the two sous-populations with and without MS according to the 7 CAD angiographic severity levels. Evidently, the number of subjects with

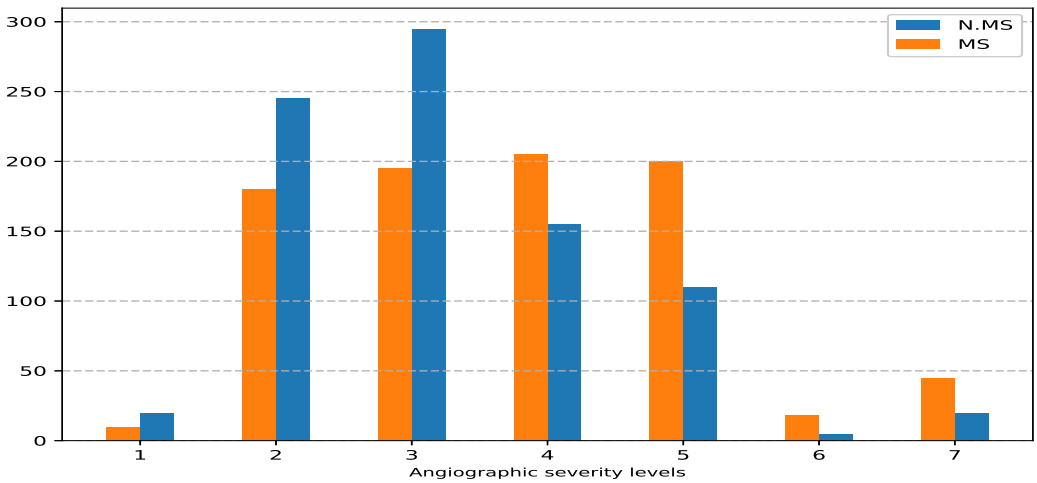


Figure 1. Distribution of the two sub-populations with and without MS according to the degree of angiographic severity.

MS is less preponderant up to level 3, but it becomes quite so from level 4. This means that the MS causes more severe coronary lesions than when there is no MS. We will perform a chi-square conformity test to confirm that this result is valid beyond the study sample. For feasibility reasons (the numbers of each theoretical class (without MS) must be greater than 5), we will group classes one and two. The result is given in Table 3.

This test is categorical, $p < 0.000000$. The MS has a positive influence on the CAD angiographic severity.

Next, we examine the influence of each individual component of MS on CAD angiographic severity. Figure 2 presents the Pearson correlation coefficients measuring the relationships between each MS component and the CAD severity variable. In the correlation matrix, values close to 1 or -1 indicate a strong and significant relationship

Table 2. Machine learning models, implementations, and optimized hyperparameters.

Model	**Implementation (Library Version)**	**Key Hyperparameters (Optimization Method)**
Logistic Regression	sklearn.linear_model.LogisticRegression (1.0.2)	penalty='l2', C=1.0 (Bayesian optimization), solver='liblinear', max_iter=1000, random_state=42
Random Forest	sklearn.ensemble.RandomForestClassifier (1.0.2)	n_estimators=200 (GridSearchCV), max_depth=12, min_samples_split=5, class_weight='balanced', random_state=42
Gaussian Naive Bayes	sklearn.naive_bayes.GaussianNB (1.0.2)	Default parameters (no tuning required)
SVM (RBF Kernel)	sklearn.svm.SVC (1.0.2)	kernel='rbf', C=10 (Optuna optimization), gamma='scale', probability=True, random_state=42
Decision Tree	sklearn.tree.DecisionTreeClassifier (1.0.2)	criterion='gini', max_depth=8 (RandomizedSearchCV), min_samples_leaf=4, random_state=42
K-Nearest Neighbors	sklearn.neighbors.KNeighborsClassifier (1.0.2)	n_neighbors=7 (Bayesian optimization), weights='distance', p=2 (Euclidean)
MLP Neural Network	sklearn.neural_network.MLPClassifier (1.0.2)	hidden_layer_sizes=(64,32), activation='relu', alpha=0.0001, learning_rate='adaptive', early_stopping=True
AdaBoost	sklearn.ensemble.AdaBoostClassifier (1.0.2)	n_estimators=150, learning_rate=0.8 (GridSearchCV), base_estimator=DecisionTree(max_depth=3)
Voting Classifier	sklearn.ensemble.VotingClassifier (1.0.2)	voting='soft', weights=[1.2, 1.1, 1.0, 0.9, 0.8] (Optimized through ensemble performance)
XGBoost	xgboost.XGBClassifier (1.5.0)	max_depth=6, learning_rate=0.01, n_estimators=500, subsample=0.8, colsample_bytree=0.7 (Optuna optimization)
LightGBM	lightgbm.LGBMClassifier (3.3.2)	num_leaves=31, max_depth=-1, learning_rate=0.05, feature_fraction=0.9 (Bayesian optimization)

Table 3. Khi2 between MS and CAD angiographic severity.

Observ.	Observ. VAR1	Theoretic. VAR2	O - T	((O - T)**2)/T
C:1	266.0000	179.3293	86.6707	41.8884
C:2	292.0000	194.8781	97.1219	48.4030
C:3	155.0000	215.6098	-60.6098	17.0379
C:4	114.0000	204.2073	-90.2073	39.8485
C:5	4.0000	15.5488	-11.5488	8.5778
C:6	19.0000	40.4268	-21.4268	11.3565
Sum	850.0000	850.0000	-.0000	167.1121

Chi²= 167.1121 dl = 5 p<0.000000.
Note: Sum of freq. Observ. & theoretic. Different.

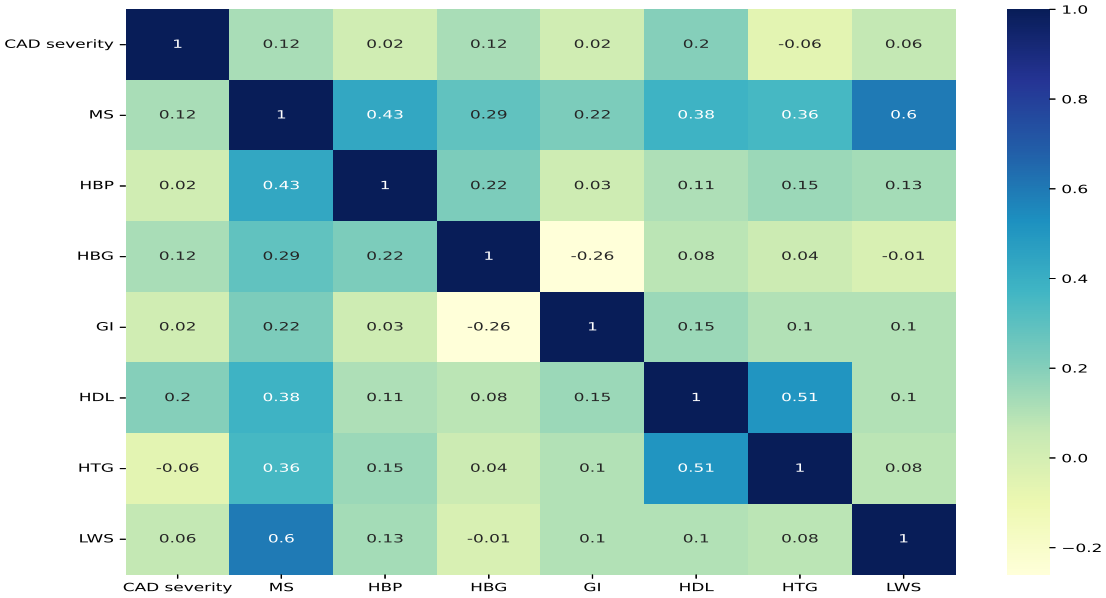


Figure 2. Matrix of correlation between determinants of CAD angiographic severity.

beyond the study sample. Conversely, a value of 0 or values near 0 suggest little to no significant correlation and should be disregarded.

In view of the results of Figure 2, we find that the MS is positively correlated with the variable1, which means that a little coronary lesion accompanies the presence of the MS because the value is only 0.12, and the maximum is 1 ". This confirms the findings of the first analysis (chi-square test). We have the other variables in the following order: Variable 6 (0.20), Variable 4 (0.12), Variable 8 (0.06). For variable 7, there is a significant correlation, but it is negative (-0.06).

The objective now is to assess the relationship between the MS components and CAD severity and to answer the following question: what are the combinations (or clusters) of MS that present the highest risk of high CAD severity? For this, odds ratios for CAD severity of different combinations of MS are used to explore the CAD severity risk of different phenotypes in triads, quartets, and quintets MS clusters by calculating the odds ratios of CAD risk. The total population is hence divided into two sub-populations:

- **The cases**: these are patients with CAD having an angiographic severity value of 3 or more.
- **The control group**: these are the patients with CAD having an angiographic severity value of 2 or less.

The so-called ”exposed” patients are those having a specified MS combination (triad, quartet or quintet MS). The ”Noexposed” patients are those without this MS combination.

As shown in Table 4, the most dangerous combination (significantly multiplies the risk of having a high severity level by 7.01) is LWS+HBP+HBG+HDL+HTG. What is also particularly striking is the fact that the HTG, as a companion in MS, played a protective role in most cases by reducing the risk of high severity by at least half (significant OR 0.42; 0.26; 0.24 and 0.14) except combinations including HTG and HDL that have increased the CAD risk of high severity (OR 2.03 and 3.51). Among the triads, the clusters with (high LWS+HBP+HBG) and (LWS+HBG +HDL) were the only combinations that significantly increased the risk of having a severe case of CAD (OR 2.31). Among the quartets, the most dangerous combination was LWS+HBP+HBG+HDL, with increased risk severity (OR 6.54). The OR for the quintet (LWS+ HBP+HDL+GI) was not significantly increased (OR 2.88, 95% CI 0.70; 11.78).

3.2. Machine learning classification of CAD severity

3.2.1. Data preparation Based on the data presented in Table 1, our collected biomedical dataset comprises 14 features. Through statistical analysis of disease severity and the results in Table 4, we developed a new nominal variable, MSC (MS Combinations), which includes 18 MS cases for patients with LWS and ischemia. The MSC variable ranges from 0 to 18, where 0 represents patients without LWS or ischemia. This adjustment reduced the number of characteristics to 8: Age, Sex, Smoking, BMI, Heredity, Clinical, Normal ECG, and MSC.

Table 4. MS Combinations versus angiographic severity.

N°	Combination	Odds ratio	Confidence Interval	Interpretation
1	LWS+HBP+HBG	2.31	(1.59 ; 3.35)	Aggravating
2	LWS+HBP+GI	1.43	(0.48 ; 4.30)	Inconclusive
3	LWS+HBP+HDL	1.91	(0.86 ; 4.26)	Inconclusive
4	LWS+HBP+HTG	0.13	(0.08 ; 0.21)	Protector
5	LWS+HBG+HDL	INCALCULABLE	significative Khi 2	Aggravating
6	LWS+HBG+HTG	0.24	(0.10 ; 0.60)	Protector
7	LWS+GI+HDL	1.61	(0.35 ; 7.31)	Inconclusive
8	LWS+GI+ HTG	0.14	(0.03 ; 0.58)	Protector
9	LWS+HDL+ HTG	0.51	(0.19 ; 1.31)	Inconclusive
10	LWS + HBP + HBG +HDL	6.54	(2.36 ; 18.08)	Aggravating
11	LWS + HBP + HBG +HTG	0.42	(0.24 ; 0.71)	Protector
12	LWS + HBP + GI +HDL	2.88	(0.70 ; 11.78)	Inconclusive
13	LWS + HBP + GI +HTG	0.26	(0.09 ; 0.71)	Protector
14	LWS + HBP + HTG +HDL	2.03	(1.07 ; 3.84)	Aggravating
15	LWS + HBG + HTG +HDL	3.51	(1.14 ; 10.75)	Aggravating
16	LWS + GI + HTG +HDL	INCALCULABLE	Khi 2 significatif	Aggravating
17	LWS+HBP+HBG+HDL+HTG	7.01	(3.20 ; 15.35)	Aggravating
18	LWS+HBP+GI+HDL+HTG	2.88	(0.70 ; 11.78)	Inconclusive

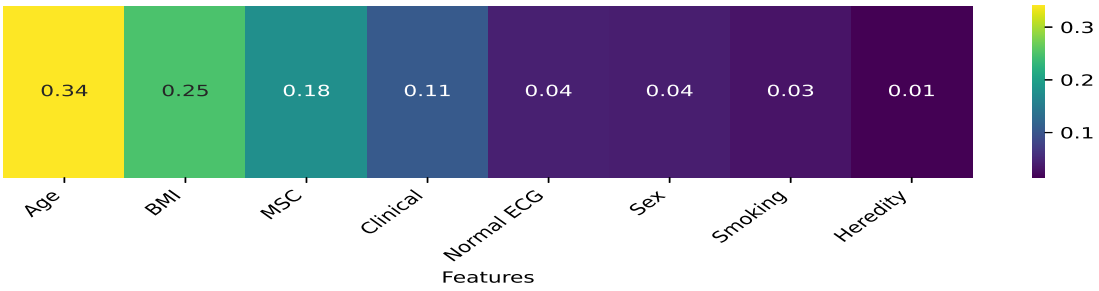


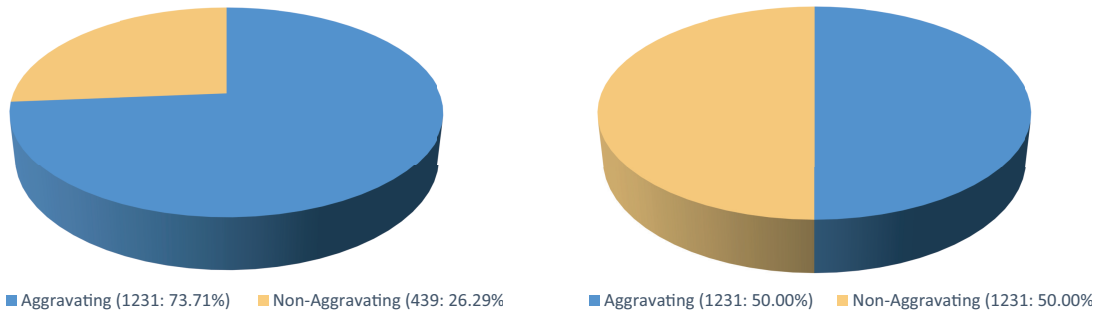
Figure 3. Mean importance scores of features with the target variable across various ML regressor models.

To enhance the accuracy and reliability of our CAD severity classification, we conducted a thorough data preparation process, including feature selection and data balancing.

The average importance of each feature was calculated using a feature importance metric, as illustrated in Figure 3. Features with importance values of 0.03 and above were considered significant contributors to the classification task, while those below 0.03 were deemed marginal contributors. Precedent literature studies in CAD and metabolic syndrome research (e.g., [24]; [46]) have used similar thresholds (0.02–0.05) to distinguish meaningful predictors from noise in clinical datasets. This range balances sensitivity (retaining weak but relevant features) and specificity (excluding spurious correlations).

Based on this threshold, several features were identified as significant. The Heredity feature, with an importance value below 0.03, was considered irrelevant and excluded from further analysis. This feature reduction aimed to enhance model performance by focusing on the most informative variables and improving computational efficiency.

Furthermore, All models were trained on SMOTE-balanced data and evaluated using 10-fold cross-validation. Prior to applying SMOTE, the dataset exhibited a skewed class distribution, which could potentially bias the model towards the majority class. By applying SMOTE, the dataset was balanced to contain an equal number of instances in each class. This process resulted in 1,231 instances per class, totaling 2,462 instances, as depicted in Figure 4. Balancing the dataset is crucial for training ML models, as it ensures that the model receives an equal representation of both classes, leading to improved generalization and classification performance.



(a) Data classes without balancing

(b) Data classes with balancing

Figure 4. The proportion of data classes with and without balancing.

3.2.2. Comparison and evaluation of machine learning models In this study, we compared the performance of various ML models in classifying the severity of CAD. The models evaluation included LR, ABT, KNN, DT, SVM, GNB, MLP, the ensemble Voting Classifier (VC), XGBoost and LightGBM. As shown in Table 5, the performance metrics used for this comparison were accuracy, precision, recall, F1-score, and specificity, averaged over a 10-fold cross-validation to ensure the robustness and generalizability of the results.

The XGBoost model achieved the highest performance across almost all metrics, with an accuracy of 0.831,

Table 5. Average performance comparison of machine learning models using 10-fold cross-validation.

ML Models	**Performance Metrics**				
	Accuracy (±std)	**Precision (±std)**	**Recall (±std)**	**F1-score (±std)**	**Specificity (±std)**
LR	0.775 (±0.033)	0.769 (±0.046)	0.794 (±0.039)	0.78 (±0.031)	0.836 (±0.034)
ABT	0.799 (±0.035)	0.792 (±0.045)	0.818 (±0.046)	0.771 (±0.037)	0.856 (±0.033)
KNN	0.738 (±0.023)	0.727 (±0.029)	0.772 (±0.036)	0.747 (±0.024)	0.80 (±0.027)
DT	0.716 (±0.036)	0.72 (±0.041)	0.718 (±0.043)	0.716 (±0.036)	0.715 (±0.034)
SVM	0.783 (±0.024)	0.776 (±0.043)	0.802 (±0.033)	0.787 (±0.024)	0.846 (±0.025)
GNB	0.768 (±0.041)	0.771 (±0.039)	0.774 (±0.051)	0.771 (±0.037)	0.816 (±0.033)
MLP	0.77 (±0.039)	0.759 (±0.042)	0.796 (±0.052)	0.716 (±0.036)	0.834 (±0.035)
VC	0.793 (±0.036)	0.788 (±0.039)	0.81 (±0.045)	0.798 (±0.034)	0.859 (±0.034)
XGBoost	**0.831 (±0.035)**	**0.844 (±0.035)**	**0.82 (±0.033)**	**0.832 (±0.034)**	**0.884 (±0.041)**
LightBGM	0.825 (±0.039)	0.839 (±0.039)	0.803 (±0.05)	0.821 (±0.051)	0.884 (±0.061)

precision of 0.848, recall of 0.82, and F1-score of 0.798. This indicates that the ensemble method effectively leverages the strengths of individual classifiers, leading to improved overall performance. The LighGBM and RF models performed well, achieving an accuracy of 0.825 and 0.0.812 respectively, with comparable precision, recall, and F1-score values, while the AdaBoost and VC models demonstrated moderate performance, with an accuracy of 0.799 and 0.793 respectively and a relatively balanced precision, recall, and F1-score.
KNN and DT had lower performance metrics across the board, with accuracies of 0.738 and 0.716, respectively. These models may not be as effective in capturing the complex patterns in the data compared to ensemble or more sophisticated individual classifiers.
Figure 5 is normalized confusion matrices for the four best-performing models: XGBoost, LightGBM, Random Forest, and AdaBoost. The four models display strong performance in identifying class 0 (milder CAD), with correct predictions ranging from 85% to 88%.

XGBoost and Random Forest outperform the others in detecting class 1 (severe CAD), achieving 81% and 80% recall respectively, with fewer false negatives. LightGBM performs slightly lower (76%), and AdaBoost shows the weakest sensitivity (75%), indicating a higher risk of missing severe cases.

Overall, the XGBoost model emerged as the most robust model for classifying CAD severity, highlighting the potential of ensemble methods in medical classification tasks. This suggests that ensemble models, by combining multiple algorithms, can achieve superior performance and should be considered in future research and applications in the medical field.

The average AUC values reported in Table 4 offer a more reliable benchmark. XGBoost and LightGBM achieved the highest AUC scores (0.884 ± 0.041 and 0.884 ± 0.061 respectively), indicating excellent discriminative ability. Random Forest also demonstrated strong performance (0.883 ± 0.032), closely followed by AdaBoost (0.856 ± 0.033) and SVM (0.846 ± 0.025). In contrast, Decision Tree presented lower AUC values, suggesting less reliable discrimination between severity classes. These results reinforce the advantage of ensemble methods in medical classification tasks, as they consistently outperformed traditional algorithms across all evaluation metrics.

To further support these findings, normalized confusion matrices for the top four models are shown in Figure 6. Each matrix is expressed in proportions (summing to 1 by row), aligning precisely with the reported accuracy and recall values. XGBoost correctly classified 85% of negative cases (non-severe CAD) and 81% of positive cases (severe CAD), showing a well-balanced sensitivity and specificity. LightGBM reached the highest true negative rate (88%) while maintaining a strong 76% true positive rate. Random Forest displayed similar performance with 85% accuracy on class 0 and 80% on class 1. AdaBoost, although slightly behind, maintained consistent performance

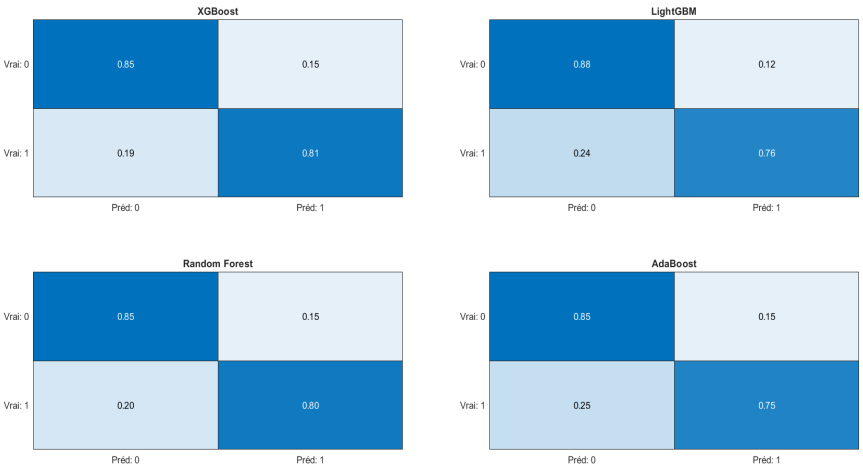


Figure 5. Confusion matrix illustrating the performance of XGBoost, LightGBM, AdaBoost and VC classifier models.

across both classes (85% and 75%). These confusion matrices confirm the robustness and generalization ability of ensemble-based models and emphasize their practical relevance in clinical CAD severity prediction.

3.2.3. Impact of MS feature on ML model performance In this part, we examine the influence of including the MS feature on the performance of some of ML models, namely (LR,ABT, KNN, DT, SVM, GNB, MLP and VC) used to classify the severity of CAD in this study. The results are illustrated in Figure 6, which compares the models' accuracy with and without the inclusion of the MS feature.

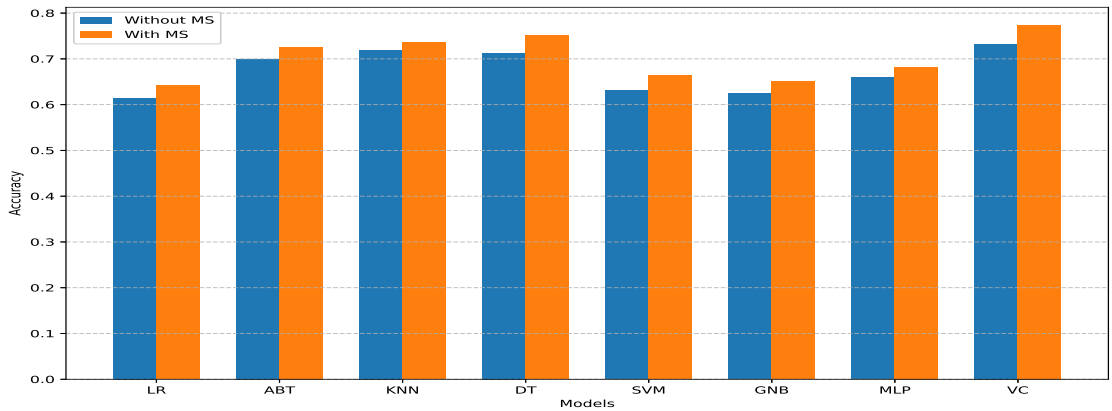


Figure 6. Comparison of ML Models Performance with and without Metabolic Syndrome Feature.

The accuracy of each model showed notable improvement when the MS feature was in- cluded. The VC model exhibited the most significant enhancement, with its accuracy increasing from 0.733 without the MS feature to 0.774 with the MS feature. Similar positive trends were observed for MLP, KNN, ABT, and DT models, confirming the clinical relevance of the MSC component.

To further highlight the inner workings of the high-performing models and improve interpretability, we applied SHAP (SHapley Additive exPlanations) analysis. This model-agnostic method provides insights into the influence

of each feature on individual predictions and global model behavior.

Table 6 shows the top SHAP-ranked features and their directional influence per model, where HDL and HBG consistently ranked as the most influential features across models, with mean absolute SHAP values of 0.22 and 0.20, respectively. This aligns with their known clinical relevance and strong association with CAD severity.
Interestingly, HTG displayed predominantly negative SHAP values in 72% of instances, suggesting a protective effect at the global level (mean SHAP = –0.08). However, stratified dependence plots revealed that HTG had a positive contribution in patients with HDL ¡ 40 mg/dL, suggesting a context-dependent role consistent with lipid interaction mechanisms.

Table 6. Top Features and Their Mean —SHAP— Values with Direction Across Models

Model	**Top Feature**	**Mean —SHAP—**	**Direction**
XGBoost	HDL	0.22	Positive
LightGBM	HBG	0.20	Positive
AdaBoost	LWS	0.18	Positive
VC	HDL	0.19	Positive

Moreover, Figure 7 is SHAP summary visualizations for the XGBoost model, representing the contribution of each feature to the prediction of CAD severity. Among all predictors, Clinical score, BMI, and MSC show the highest mean absolute SHAP values, showing their strong influence on the model's output. Notably, higher values of MSC are associated with increased SHAP values, which confirm that patients with multiple metabolic syndrome components are more likely to be classified as CAD-aggravated. Conversely, features such as Smoking and Heredity have relatively lower SHAP contributions, suggesting limited predictive weight in the context of the trained model.

This highlights the importance of incorporating relevant clinical features such as MS—particularly HDL and HBG—in CAD severity prediction and in enhancing the predictive capabilities of ML models in medical applications.

3.2.4. Gender-Specific Model Performance Based on the balanced training dataset, the XGBoost classifier's gender-specific performance was analyzed, stratifying predictions to evaluate fairness and generalizability (Table 7).

The XGBoost model demonstrated robust performance across genders, with comparable AUC scores for men (0.86) and women (0.85). While women exhibited marginally lower accuracy ($\Delta = 0.02$) and F1 scores ($\Delta = 0.02$), these differences were not statistically significant ($p > 0.05$, Wilcoxon signed-rank test), suggesting equitable predictive behavior. The balanced resampling strategy preserved fairness, with equalized odds ratios close to 1.0 (F1 men/women ratio = 1.03).

Table 7. Performance of XGBoost Model Stratified by Gender

Subgroup	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Overall	0.83	0.84	0.82	0.83	0.88
Men	0.81	0.83	0.79	0.81	0.86
Women	0.79	0.80	0.78	0.79	0.85

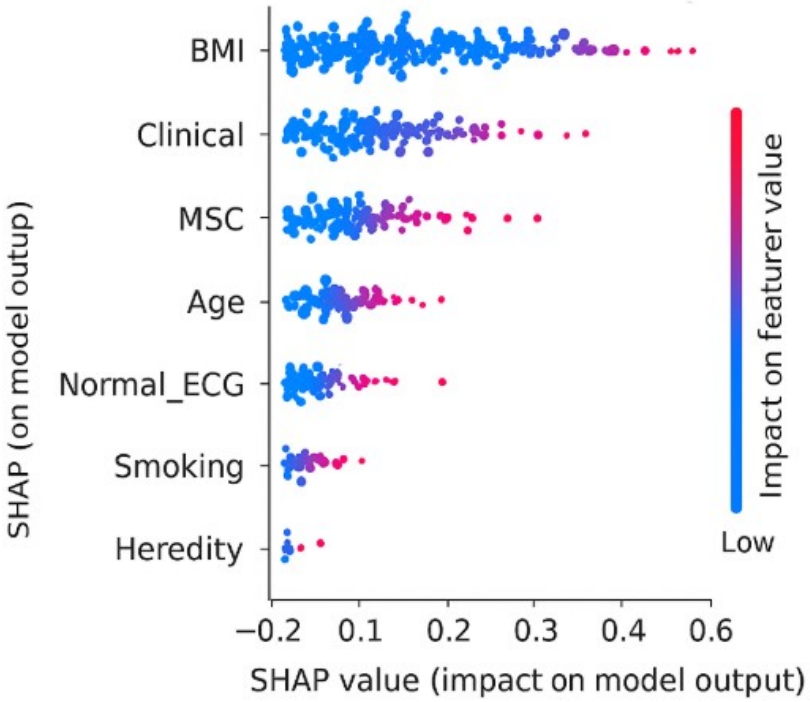


Figure 7. Top Contributing Features Identified by SHAP Analysis Across XGBoost Model.

4. Discussion and conclusion

In view of the obtained results, several significant findings have emerged:

- Out of three women with CAD, two will also have MS.
- CAD particularly affects men. However, among patients with CAD, it is mostly women who are predisposed to present MS.
- The presence of MS causes more severe CAD damage than its absence.
- HDL is the most aggravating factor for coronary lesions, followed by HBG and LWS. We cannot draw any conclusions for the HBP and GI variables as their correlations are not significant at the 5% level. Rather, the HTG variable has a protective influence.
- Compared to the work of Kim et al. [24], where MS was diagnosed for those having three or more of five components, our study includes GI as an additional component. Therefore, those having an excess of waist size plus two or more of the five remaining components are classified as having MS.
- On the predictive ability of MS and its components for CAD, the study by Kim et al. [24] resolved that HBG was the only predictive factor for CAD. At the same time, the results for HDL, LWS, HBP, and HTG were inconclusive. In our study, the most associated factors with CAD severity were LWS, HBG, HDL, and HTG. The results for GI and HBP are inconclusive.
- Kim et al. [24] also showed that the odds ratios (ORs) of the 11 possible combinations of individual MS components varied widely as CAD risk factors. The MS cluster, which included HBG, HBP, and HDL, was associated with the highest CAD risk. In our study, the combination of LWS, HBP, HBG, HDL, and HTG was associated with the highest CAD risk and is the most appropriate predictive model for CAD angiographic severity.

In light of the results obtained from our study on ML models for classifying the severity of CAD, several key insights can be drawn that complement the preliminary discussion on the relationships between MS and CAD:

- The eXtreme Gradient Boosting (XGBoost) emerged as the most effective model, achieving the highest performance metrics across accuracy, precision, recall, and F1-score. This finding underscores the potential of ensemble methods in medical classification tasks, particularly in complex scenarios like CAD, where multiple factors interact.
- Moreover, the results indicate that the ML models, particularly the XGBoost, can be instrumental in identifying the most significant predictors of CAD severity, such as LWS, HBG, HDL, and HTG. This aligns with our earlier findings that HDL is a critical factor in aggravating coronary lesions. The ability of the XGBoost model to integrate these variables may provide a more nuanced understanding of how MS components interact with CAD, enhancing predictive accuracy.
- The confusion matrices for the top-performing models (XGBoost, LightGBM, ABT, and VC) further elucidate the classification performance, revealing the strengths and weaknesses of each model in identifying "Aggravating" versus "Non-Aggravating" cases.
- Our stratified analysis doesn't reveal any significant gender-based performance disparities. However, the slight numerical differences (e.g., recall for women: 0.78 vs. men: 0.79) suggest further investigation into sex-specific risk by integrating fairness-aware algorithms (e.g., adversarial debiasing) to explicitly optimize for equity.
- Our results confirm the "triglyceride paradox" emphasizing the complex dual role of lipid metabolism in heart disease. It is well-known that elevated triglyceride levels (HTG) are contributors to early plaque formation in arteries [44], however emerging evidence shows that they might offer protective benefits in patients with established (CAD). This counterintuitive phenomenon seems related to three main mechanisms: reducing inflammatory responses [5], preventing toxic fat buildup in cells [16], and stabilizing vulnerable plaque areas [5].
- Paradoxically increase of coronary artery disease risk in the case of the coexistence of (HTG) and low HDL cholesterol (HDL-C) is due to elevated triglycerides that promote cholesteryl ester transfer via CETP, producing dysfunctional HDL particles with impaired cholesterol efflux capacity. This pro-inflammatory interaction exacerbates oxidative stress and endothelial dysfunction, negating the potential protective effects of isolated HTG. Therapeutic strategies combining fibrates and omega-3 fatty acids, which target both triglycerides and HDL functionality, have shown greater cardiovascular benefits, highlighting the importance of personalized lipid risk assessment [34] and [12].

In conclusion, the integration of ML models in assessing CAD severity not only complements the findings regarding the impact of MS but also emphasizes the need for further research in this area. Future studies could explore applying these models to larger datasets and different populations, potentially leading to more robust predictive tools to inform clinical decision-making. Additionally, investigating the interplay between the various components of MS and their collective impact on CAD severity through ML could yield valuable insights that enhance understanding and treatment strategies in cardiovascular health.
Regrading practical implications, based on our results, several actionable insights can be derived to guide healthcare providers in clinical decision-making for the management of (CAD) severity in patients with metabolic syndrome (MS). As such, we recommend that clinicians:

- Prioritize early screening and intervention in patients presenting with these high-risk MS profiles.

- Implement aggressive lifestyle modification strategies (e.g., weight management, dietary control) in overweight individuals with MS.

- Pay closer attention to ECG abnormalities, even when other clinical signs are subtle, as they were frequently present in more severe CAD cases.

- Investigate familial history systematically, as heredity emerged as a key predictive factor of severity.

These findings suggest a need for personalized risk stratification based on the specific MS profile, rather than a binary MS presence/absence.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research received no specific grant funding from agencies in the public, commercial or not-for-profit sectors.

REFERENCES

1. A. Ainiwaer, W. Q. Hou, K. Kadier, R. Rehemuding, P. F. Liu, H. Maimaiti, L. Qin, X. Ma, and J. G. Dai, *A machine learning framework for diagnosing and predicting the severity of coronary artery disease*, Reviews in Cardiovascular Medicine, vol. 24, no. 6, pp. 168, 2023.
2. A. Akella and S. Akella, *Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution*, Future Science OA, vol. 7, no. 6, 2021.
3. L. B. Almeida, *Multilayer perceptrons*, Handbook of Neural Computation, pp. C1–2, CRC Press, 2020.
4. F. Anctil, C. Michel, C. Perrin, and V. Andréassian, *A soil moisture index as an auxiliary ANN input for stream flow forecasting*, Journal of Hydrology, vol. 286, no. 1–4, pp. 155–167, Jan. 2004.
5. A. Arnold, *Primary hyperparathyroidism: molecular genetic insights and clinical implications*, Endocrine abstracts, vol. 50, 2017.
6. O. Y. Atkov, S. G. Gorokhova, A. G. Sboev, E. V. Generozov, E. V. Muraseyeva, S. Y. Moroshkina, and N. N. Cherniy, *Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters*, Journal of Cardiology, vol. 59, no. 2, pp. 190–194, Mar. 2012.
7. H. Ayatollahi, L. Gholamhosseini, and M. Salehi, *Predicting coronary artery disease: a comparison between two data mining algorithms*, BMC Public Health, vol. 19, no. 1, Apr. 2019.
8. W. G. Baxt and H. White, *Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction*, Neural Computation, vol. 7, no. 3, pp. 624–638, 1995.
9. R. Bhuvaneswari and K. Kalaiselvi, *Naive Bayesian classification approach in healthcare applications*, International Journal of Computer Science and Telecommunications, vol. 3, no. 1, pp. 106–112, 2012.
10. Y. Cao, Q.-G. Miao, J.-C. Liu, and L. Gao, *Advance and Prospects of AdaBoost Algorithm*, Acta Automatica Sinica, vol. 39, no. 6, pp. 745–758, Jun. 2013.
11. B. Chamontin, *Hypertension artérielle de l'adulte. Epidémiologie, étiologie, physiopathologie, diagnostic, évolution, pronostic traitement de l'hypertension artérielle essentielle*, La Revue du praticien (Paris), vol. 51, no. 15, pp. 1697–1713, 2001.
12. P. Cullen, *Evidence that triglycerides are an independent coronary heart disease risk factor*, The American Journal of Cardiology, vol. 86, no. 9, pp. 943–949, 2000.
13. X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, *A survey on ensemble learning*, Frontiers of Computer Science, vol. 14, no. 2, pp. 241–258, Aug. 2019.
14. C. Eyupoglu and O. Karakuş, *Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques*, Journal of Clinical Medicine, vol. 13, no. 10, pp. 2868, May 2024.
15. A. Ghachem, *Obésité, facteurs de risque et complications cardiométaboliques chez les personnes âgées de 50 ans et plus: Mieux comprendre pour mieux intervenir*, PhD Thesis, Université de Sherbrooke, Avril 2018. Available at https://savoirs.usherbrooke.ca/bitstream/handle/11143/5911/Ghachem_Ahmed_PhD_2018.pdf
16. I. J. Goldberg, R. H. Eckel, and R. McPherson, *Triglycerides and heart disease: still a hypothesis?*, Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 31, no. 8, pp. 1716–1725, 2011.
17. M. Grandini, E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: an Overview*, arXiv, 2020, https://arxiv.org/abs/2008.05756
18. I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003.
19. M. Hermans, S. Ahn, and M. Rousseau, *Traitement des dyslipidémies et risque de cardiopathie ischémique dans le diabète: méta-analyse de 198.930 patients inclus dans 57 études randomisées*, Louvain médical, vol. 134, no. 6, pp. 289, 2015.
20. Y. C. N. Houehanou Sonou, *Épidémiologie des facteurs de risque cardiovasculaire en population tropicale-cas du Bénin*, PhD Thesis, Limoges, December 2015.
21. A. A. Huang and S. Y. Huang, *Use of machine learning to identify risk factors for coronary artery disease*, PloS one, vol. 18, no. 4, pp. e0284103, 2023.
22. H. Kaya, F. Ertaş, M. Oylumlu, M. Z. Bilik, A. Yıldız, M. Yüksel, N. Polat, H. Acet, F. Işık, and M. S. Ülgen, *Relation of epicardial fat thickness and brachial flow-mediated vasodilation with coronary artery disease*, Journal of Cardiology, vol. 62, no. 6, pp. 343–347, Dec. 2013.
23. N. Khaltaev and S. Axelrod, *Countrywide cardiovascular disease prevention and control in 49 countries with different socio-economic status*, Chronic Diseases and Translational Medicine, vol. 8, no. 4, pp. 296–304, 2022.
24. J.-Y. Kim, H.-S. Mun, B. K. Lee, S. B. Yoon, E.-Y. Choi, P.-K. Min, Y.-W. Yoon, B.-K. Hong, S.-J. Rim, and H. M. Kwon, *Impact of Metabolic Syndrome and Its Individual Components on the Presence and Severity of Angiographic Coronary Artery Disease*, Yonsei Medical Journal, vol. 51, no. 5, pp. 676, 2010.

25. Y. B. Koffi, *Modélisation pluie-débit en région tropicale humide : application des réseaux de neurones sur quatre stations hydrométriques du Bandama Blanc (Bada, Marabadiassa, Tortiya et Bou) situées au Nord de la Côte d'Ivoire. Thèse de l'Université*, Physio-Géo, vol. 3, pp. 1–3, Jan. 2009.
26. B. Kolukisa and B. Bakir-Gungor, *Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis*, Computer Standards & Interfaces, vol. 84, pp. 103706, Mar. 2023.
27. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, *Machine learning applications in cancer prognosis and prediction*, Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.
28. O. Kramer, *Scikit-Learn*, in Machine Learning for Evolution Strategies, Springer International Publishing, Cham, 2016, pp. 45–53.
29. S. Kumari, D. Kumar, and M. Mittal, *An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier*, International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40–46, Jun. 2021.
30. P. A. McCullough, *Coronary Artery Disease*, Clinical Journal of the American Society of Nephrology, vol. 2, no. 3, pp. 611–616, May 2007.
31. S. Nusinovici, Y. C. Tham, M. Y. C. Yan, D. S. W. Ting, J. Li, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng, *Logistic regression was as good as machine learning for predicting major chronic diseases*, Journal of Clinical Epidemiology, vol. 122, pp. 56–69, Jun. 2020.
32. S. Patil, J. W. Henry, M. Rubenfire, and P. D. Stein, *Neural Network in the Clinical Diagnosis of Acute Pulmonary Embolism*, Chest, vol. 104, no. 6, pp. 1685–1689, Dec. 1993.
33. M. Pellegrin, L. Mazzolai, A. Berthelot, and P. Laurant, *Dysfonction endothéliale et risque cardiovasculaire. L'exercice protège la fonction endothéliale et prévient la maladie cardiovasculaire*, Science & Sports, vol. 24, no. 2, pp. 63–73, Apr. 2009.
34. X. Peng, Z. Lian, V. O'Brien, J. Xiao, B. A. Litchfield, X.-Y. Dai Perrard, L. Xu, J. Ni, A. Mukherjee, T. Simmons, et al., *Foamy monocytes and atherogenesis in mice with combined hyperlipidemia and effects of antisense knockdown of apoCIII*, Journal of Lipid Research, vol. 66, no. 4, 2025.
35. G. M. Reaven, *Role of Insulin Resistance in Human Disease*, Diabetes, vol. 37, no. 12, pp. 1595–1607, Dec. 1988.
36. N. Revathi, P. M. Kavitha, D. Narayani, S. Irin Sherly, M. Robinson Joel, and P. Jose, *Heart Disease Prediction using Multimodal Data with Multi-Layer Perceptron*, International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 1728–1737, Jun. 2024.
37. M. Shoaib, A. Junaid, G. Husnain, M. Qadir, Y. Y. Ghadi, S. S. Askar, and M. Abouhawwash, *Advanced detection of coronary artery disease via deep learning analysis of plasma cytokine data*, Frontiers in Cardiovascular Medicine, vol. 11, pp. 1365481, 2024.
38. A. Sirajuddin, S. M. Mirmomen, S. J. Kligerman, D. W. Groves, A. P. Burke, F. Kureshi, C. S. White, and A. E. Arai, *Ischemic Heart Disease: Noninvasive Imaging Techniques and Findings*, RadioGraphics, vol. 41, no. 3, pp. 200125, May 2021.
39. S. Suthaharan, *Decision Tree Learning*, in Integrated Series in Information Systems, Springer US, 2016, pp. 237–269.
40. S. Suthaharan, *Support Vector Machine*, in Integrated Series in Information Systems, Springer US, 2016, pp. 207–235.
41. N. Tasnim, S. A. Al Mamun, M. S. Islam, M. S. Kaiser, and M. Mahmud, *Explainable Mortality Prediction Model for Congestive Heart Failure with Nature-Based Feature Selection Method*, Applied Sciences, vol. 13, no. 10, pp. 6138, May 2023.
42. K. A. Wilmot, M. O'Flaherty, S. Capewell, E. S. Ford, and V. Vaccarino, *Coronary Heart Disease Mortality Declines in the United States From 1979 Through 2011: Evidence for Stagnation in Young Adults, Especially Women*, Circulation, vol. 132, no. 11, pp. 997–1002, Sep. 2015.
43. W. SMAILI, *Profil épidémiologique, clinique et thérapeutique du syndrome métabolique au sein du service de GHE*, PhD Thesis, Cadi ayyad university, September 2017. Available at http://wd.fmpm.uca.ma/biblio/theses/annee-htm/FT/2017/these178-17.pdf
44. T.-l. Xia, Y.-m. Li, F.-y. Huang, H. Chai, B.-t. Huang, Q. Li, Z.-g. Zhao, Y.-b. Liao, Z.-l. Zuo, Y. Peng, et al., *The triglyceride paradox in the mortality of coronary artery disease*, Lipids in Health and Disease, vol. 18, pp. 1–7, 2019.
45. W. Yu, L. Yang, F. Zhang, B. Liu, Y. Shi, J. Wang, X. Shao, Y. Chen, X. Yang, and Y. Wang, *Machine learning to predict hemodynamically significant CAD based on traditional risk factors, coronary artery calcium and epicardial fat volume*, Journal of Nuclear Cardiology, vol. 30, no. 6, pp. 2593–2606, Dec. 2023.
46. T. Zhang, S. Huang, P. Xie, X. Li, Y. Pan, Y. Xu, P. Han, F. Ding, J. Zhao, and H. Tang, *Development of Machine Learning Tools for Predicting Coronary Artery Disease in the Chinese Population*, Disease Markers, vol. 2022, no. 1, pp. 6030254, 2022.
47. Z. Zhang, *Introduction to machine learning: k-nearest neighbors*, Annals of Translational Medicine, vol. 4, no. 11, pp. 218–218, Jun. 2016.