sensors

MDPI

Article

Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models

Maria Trigka * and Elias Dritsas

Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece
* Correspondence: trigka@ceid.upatras.gr

Abstract: The heart is the most vital organ of the human body; thus, its improper functioning has a significant impact on human life. Coronary artery disease (CAD) is a disease of the coronary arteries through which the heart is nourished and oxygenated. It is due to the formation of atherosclerotic plaques on the wall of the epicardial coronary arteries, resulting in the narrowing of their lumen and the obstruction of blood flow through them. Coronary artery disease can be delayed or even prevented with lifestyle changes and medical intervention. Long-term risk prediction of coronary artery disease will be the area of interest in this work. In this specific research paper, we experimented with various machine learning (ML) models after the use or non-use of the synthetic minority oversampling technique (SMOTE), evaluating and comparing them in terms of accuracy, precision, recall and an area under the curve (AUC). The results showed that the stacking ensemble model after the SMOTE with 10-fold cross-validation prevailed over the other models, achieving an accuracy of 90.9 %, a precision of 96.7%, a recall of 87.6% and an AUC equal to 96.1%.

Keywords: healthcare; long-term risk prediction; machine learning; coronary artery disease; feature analysis

Citation: Trigka, M.; Dritsas, E. Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models. *Sensors* **2023**, *23*, 1193. https://doi.org/10.3390/s23031193

Academic Editor: Wan-Young Chung

Received: 28 December 2022
Revised: 17 January 2023
Accepted: 18 January 2023
Published: 20 January 2023

1. Introduction

The heart is a tireless muscular pump, the size of a large fist and weighing 300–400 g. It circulates tons of blood during human life. Cardiovascular disease remains the leading cause of death despite significant advances in medical science. It needs special attention and awareness to minimize the factors that cause it, as nowadays, the habits and lifestyle of modern people directly impact it [1,2].

The coronary arteries are the arteries that transport blood to the heart muscle and supply it with the necessary ingredients for its function. The term "coronary artery disease" is used to describe the narrowing of these arteries, which is caused by the accumulation of atherosclerotic material in their lumen. Due to the stenosis, the heart muscle is not adequately supplied with blood–especially in situations where it has increased needs–and this causes myocardial ischemia [3]. In the vast majority of cases, CAD is caused by the progressive accumulation of atherosclerotic material, which narrows the lumen of the arteries and causes myocardial ischemia. Atherosclerotic material is a soft, fatty material that forms on the inner surface of the arteries by interacting with blood elements (cells and coagulation factors) and fats carried by the blood. Atherosclerotic plaque "hardens" over the years due to calcium deposition [4].

Angina pain is a common manifestation of insufficient perfusion of the myocardium, and manifests in discomfort in the centre of the chest which may be tight, or feel like burning or pressure. Angina may be felt in both hands, in the area of the neck, lower jaw, in the mid-shoulder area and in the epigastrium. Sometimes, when the pain is intense, sweat, nausea or vomiting occur. Manifestations of CAD include [4,5]:

- **The asymptomatic period**: The process of atherosclerosis does not cause symptoms. Furthermore, patients who do not have severe coronary artery stenosis may have no symptoms, despite the presence of atherosclerotic lesions in the coronary arteries [6].
- **Stable angina**: The appearance of angina pain either during physical activity or during intense emotional stress. Stable angina is generally a relatively benign clinical condition and usually offers the opportunity to select and apply the appropriate treatment [7].
- **Unstable angina**: The appearance of angina pain at rest. This is a more dangerous form of coronary artery disease, which is why it has been described as pre-infarction angina. It is clear that such an unstable condition must be treated with hospitalization so that the administration of appropriate treatment can be commenced in order to avoid undesirable progression to myocardial infarction [8].
- **Acute myocardial infarction**: This is the necrosis of an area of the heart muscle that manifests itself with typical angina, which, however, is prolonged, does not stop with rest and lasts more than half an hour. The immediate transfer of the patient to a hospital is imperative, because only in a specialized area and by specialized personnel can such a serious medical problem be treated with the greatest possible rate of success [9].
- **Sudden cardiac death**: This is the most dramatic manifestation of the entire clinical spectrum of coronary artery disease [10].

Coronary artery disease is mainly due to atherosclerosis of the coronary arteries. The cause of atherosclerosis is not singular; for it to occur, many factors work together, i.e., it is a multifactorial disease. The factors that all act together as the cause of atherosclerosis are called risk factors or predisposing factors and are the following: gender, age, heredity, hypercholesterolemia, smoking, hypertension, obesity and sedentary lifestyle, diabetes mellitus, metabolic syndrome, chronic renal failure and stress [11,12].

The prevention of heart disease and, therefore CAD lies mainly in changing lifestyles and adopting healthier habits. A balanced diet, exercise and getting rid of bad habits will keep the arteries strong and clean of atherosclerotic plaques. More specifically, some ways to help improve the health of the cardiovascular system are the following: smoking cessation, regular physical exercise, control of blood pressure, low cholesterol and lack of diabetes, maintaining a stable body weight, reducing stress, eating a Mediterranean diet rich in fruits and vegetables, avoiding salt, consuming of foods rich in fibre and limiting alcohol consumption [13,14].

Nowadays, medicine has a variety of modern diagnostic tests, which, in cooperation with Information technology and, especially, the fields of artificial intelligence (AI) and machine learning (ML), in the hands of cardiologists are powerful weapons for the prevention or diagnosis of coronary artery disease. ML techniques now play an important role in the early prediction of disease complications in diabetes (as classification [15,16] or regression tasks for continuous glucose prediction [17,18]), cholesterol [19,20], hypertension [21,22], chronic obstructive pulmonary disease (COPD) [23], COVID-19 [24], stroke [25], chronic kidney disease (CKD) [26], liver disease (LD) [27], sleep disorders [28,29], hepatitis C [30], cardiovascular diseases (CVDs) [31], lung cancer [32], and metabolic syndrome [33] etc. In particular, the long-term risk prediction of CAD will concern us in the context of this study. The main contributions of the present research work are the following:

- Data preprocessing is achieved with the SMOTE. In this way, the instances of the dataset are distributed in a balanced way, allowing us to design robust classification models to ensure a highly accurate prediction of CAD occurrence.
- Features' importance evaluation is performed considering two commonly used methods, the gain ratio and random forest methods. This analysis is made using the initial unbalanced data and those obtained after class balancing using SMOTE.
- Experimental evaluation is performed with various ML models, after the use or not of SMOTE, evaluating and comparing them in terms of accuracy, precision, recall and AUC. The experimental results indicated that the stacking ensemble model after

SMOTE, with 10-fold cross-validation, prevailed over the other ones, constituting the main proposition of this research paper.

The rest of the paper is organized as follows. In Section 2, a dataset description and analysis of the methodology followed are made. Additionally, in Section 3, we discuss the acquired research results. Then, Section 4 discusses the relevant works with the subject under consideration. Finally, conclusions and future directions are outlined in Section 5.

2. Materials and Methods

In this section, an overview of the dataset we relied on is carried out, the methodology followed is captured, details of the experimental setup are noted, and brief descriptions of the ML models we experimented with and their evaluation metrics are outlined.

2.1. Dataset Description

In this research paper, we used a publicly available dataset [34]. The present dataset includes 3655 instances. It has 15 features, 7 of which are nominal and 8 numerical. Specifically nominal are gender [35], education [36], current smoker [37], blood pressure medication (BPMeds) [38], prevalent stroke (prevStroke) [39], prevalent hypertension (prevHyp) [40] and diabetes [41], while numerical are age [9], cigarettes per day (cigs per day) [42], total cholesterol (totChol) [43], systolic blood pressure (sysBP) [44], diastolic blood pressure (diaBP) [45], body mass index (BMI) [46], heart rate [47] and glucose [48]. The target class, denoted as CAD, is binary and refers to coronary artery disease occurrence or not.

Further statistical details about the features in terms of the target class labels are presented in Table 1. More specifically, the number of participants who have been diagnosed with CAD is 556 (15.2%). Furthermore, the number of women is 2033 (55.6%), while the number of men is 1622 (44.4%). The age of the participants varies from 32 to 70 years.

Table 1. Numerical and nominal features' description in the initial dataset before SMOTE.

Attribute	Description			Attribute	Description
	Min	Max	Mean ± stdDev	**Gender**	male (1622), Female (2033)
Age	32	70	49.5 ± 8.56	**Education**	PhD (423), BSc (1100), High School (1526), MSc (606)
Cigs/day	0	70	9 ± 11.92		
totChol	113	464	236.8 ± 43.69	**Current smoker**	Yes (1788), No (1867)
SysBP	83.5	295	132.3 ± 22.1		
DiaBP	48	142.5	82.9 ± 11.97	**BPMeds**	Yes (111), No (3544)
BMI	15.54	56.8	25.8 ± 4.07	**prevStroke**	Yes (21), No (3634)
Heart rate	44	143	75.7 ± 11.99	**prevHyp**	Yes (1138), No (2517)
Glucose	40	394	81.8 ± 23.89	**Diabetes**	Yes (98), No (3557)

2.2. Methodology

The following subsections emphasize the methodology followed in order to evaluate the ML models we experimented with.

2.2.1. CAD Risk Prediction

The long-term risk prediction of coronary artery disease is formulated as a classification problem with two possible classes c = "CAD" or c = "non-CAD". The trained ML models will be able to predict the class of a new unclassified instance either as CAD or non-CAD, based on the input features' values, and thus predict the risk of coronary artery disease.

2.2.2. Data Preprocessing

The accurate identification of CAD and non-CAD instances may be impacted by the unbalanced distribution of the instances in the two classes. Here, an oversampling method is applied, namely SMOTE [49], which is based on the K-Nearest Neighbors (KNN) [50] classifier with $K = 5$ and creates synthetic data [51] on the minority class (see Algorithm 1). The instances in the CAD class are oversampled, such that the subjects in the two classes are uniformly distributed. After the application of SMOTE, the dataset becomes balanced, the number of participants is 6198 and the class variable includes 3099 CAD and 3099 non-CAD instances.

Algorithm 1 SMOTE

Input: M (number of samples in the minority class), N (% ratio of synthetic minority samples for class balancing), K (number of nearest neighbors), s_{syn} synthetic instance;
Choose randomly a subset $\mathcal{S}$ of the minority class data of size $S = \frac{N}{100}M$ (synthetic samples in the minority class) such that the class labels are uniformly distributed;
for all $s_i \in \mathcal{S}$ **do**
(1) Find the K nearest neighbors;
(2) Randomly select one of KNN, called $\hat{s}_i$;
(3) Calculate the distance $d_{i,k} = \hat{s}_i - s_i$ between the randomly selected NN $\hat{s}_i$ and the instance s_i;
(4) The new synthetic instance is generated as $s_{syn} = s_i + \delta d_{i,k}$ (where $\delta = rand(0, 1)$ is a random number between 0 and 1);
end for
Repeat steps number 2–4 until the desired proportion of minority class is met.

The number of women is 2805 (45.3%), while the number of men is 3393 (54.7%). Finally, statistical details about the features in the balanced data are outlined in Table 2.

Table 2. Numerical and nominal features' description after SMOTE.

Attribute	Description			Attribute	Description
	Min	Max	Mean ± stdDev	**Gender**	Male (3393), Female (2805)
Age	32	70	51.5 ± 8.34	**Education**	Phd (665), BSc (1693), High School (3198), MSc (642)
Cigs/day	0	70	9.4 ± 11.79		
totChol	113	464	240.5 ± 44.18	**Current smoker**	Yes (2803), No (3395)
sysBP	83.5	295	136.8 ± 23.8		
diaBP	48	142.5	84.7 ± 12.59	**BPMeds**	Yes (111), No (6087)
BMI	15.54	56.8	26 ± 3.91	**prevStroke**	Yes (21), No (6177)
Heart rate	44	143	75.8 ± 11.45	**prevHyp**	Yes (2335), No (3863)
Glucose	40	394	84.3 ± 30.95	**Diabetes**	Yes (183), No (6015)

2.2.3. Features Analysis

In the context of this subsection, our aim is to investigate the importance of the features that represent the instances of the dataset. Two different methods were used: gain ratio and random forest.

First, we employed the gain ratio (GR) method [52] to measure the importance of the features in predicting the target class, calculating it as $GR(X_i) = \frac{H(C) - H(C|X_i)}{H(X_i)}$, for $i = 1, 2, \ldots, 15$. In the previous equation, the denominator is the entropy of feature X_i defined as $H(X_i) = -\sum_{x_i \in V_i} p(x_i) log_2(p(x_i))$ (with V_i be the set of different values and p_{x_i} denotes the probability of state x_i of feature X_i). Furthermore, the left term in the nominator is the entropy of class variable C defined as $H(C) = -\sum_{c \in C} p(c) log_2(p(c))$ (with $p(c)$ being the probability of state $c \in C = \{CAD, Non - CAD\}$). Finally, the right term in

the nominator is the conditional entropy of feature X_i given the C which is calculated as $H(C|X_i) = -\sum_{c \in C} \sum_{x_i \in V_i} p(c|x_i) log_2(p(c|x_i))$ (where $p(c|x_i)$ is the related conditional probability of state c given value x_i).

In Figure 1, we exploited the GR method to capture the features' order of importance before and after the use of SMOTE. We observed that after SMOTE, heart rate and cigarettes per day were categorized third and fourth in order, respectively, which without SMOTE were last in order with zero scores.

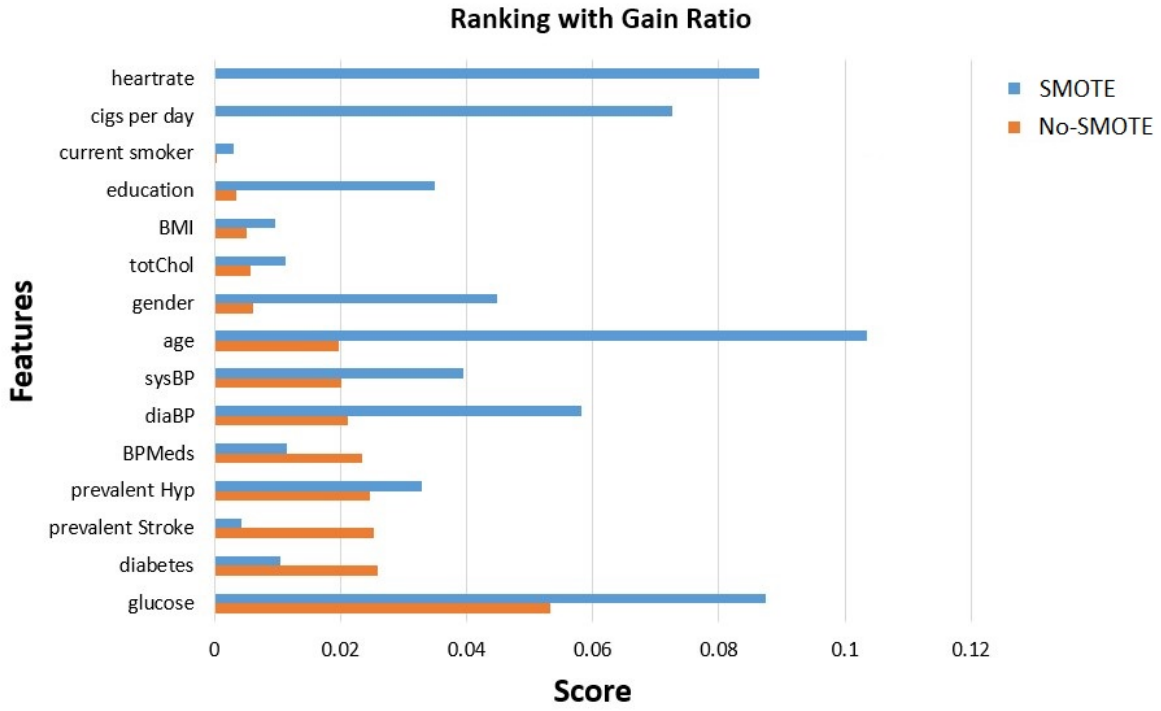

Figure 1. Gain ratio features' importance evaluation before and after SMOTE.

Random forest is a popular machine-learning algorithm characterized by high-accuracy predictive ability, low overfitting (better generalization), and easy interpretability. Feature selection using random forest is categorized as an embedded method that achieves a ranking of importance by the Gini impurity index. Gini impurity is computed at every node split during the construction of a decision tree and measures the quality of the split in terms of separating the samples of the different classes in the specific node. The higher the increment in leaf purity, the higher the importance of the feature. This is applied for each tree and averaged among all the trees normalized to 1. So, the sum of the importance scores calculated by a random Forest is 1. Gini impurity index is computed based on Equation [53]:

$$G = \sum_{i=1}^{c} p_i (1 - p_i)^2, \quad (1)$$

where c denotes the number of classes and p_i is the probability of a sample being categorized in class i.

In Figure 2, features' importance is computed based on random forest, which exploits (1). Observing this figure, the features' importance was essentially increased, and some of them, such as BMI, cigarettes per day, and heart rate were elevated from the bottom to the top of the hierarchy. Both in the case of random forest, most of the features' importance was enhanced except for diabetes, stroke prevalence and blood pressure medication (BPMeds). For the models' training and testing, all of these features were exploited.

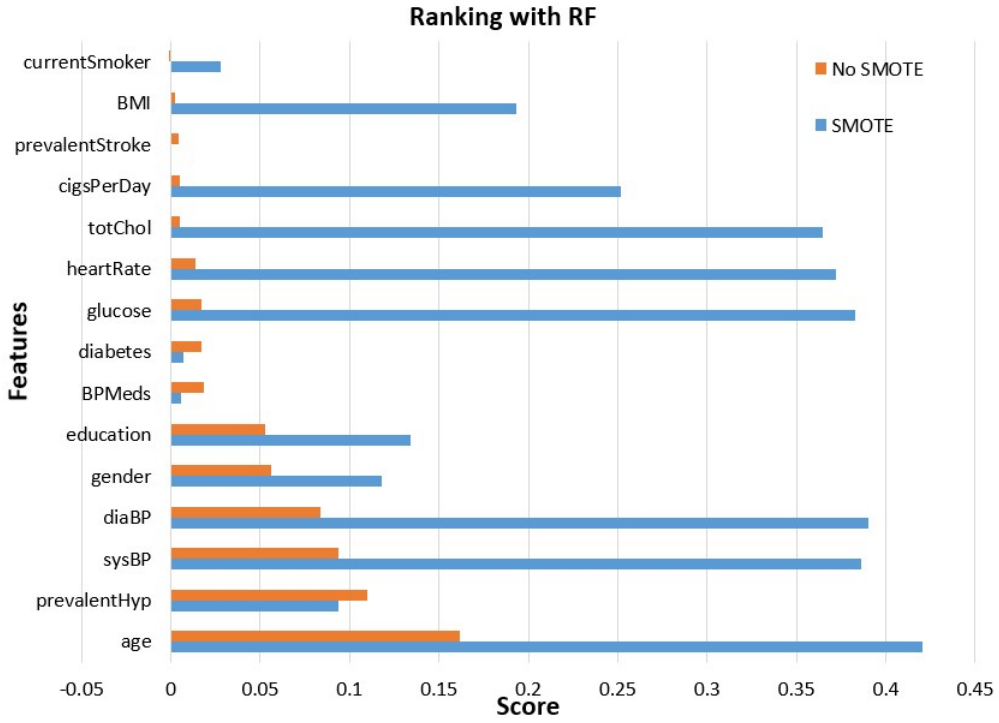

Figure 2. RF features' importance evaluation before and after SMOTE.

2.3. Machine Learning Models

In this research article, we experimented with various ML models to discover which one outperforms the others by evaluating their prediction performance. Specifically, we focused on naive Bayes (NB) [54], which assigns an instance to that class for which the conditional probability of the features' set given class label is maximized, and logistic regression (LR) [55], which are probabilistic classifiers. Furthermore, we used a decision-tree-based model, especially J48 [56]. From ensemble ML algorithms, bagging [57], random forest (RF) [58], rotation forest (RotF) [59], voting [60] and stacking [61] were exploited. Furthermore, a fully connected class of feedforward artificial neural network (ANN), i.e., multilayer perceptron (MLP) [62], and KNN, a distance-based classifier, were evaluated. Finally, in Table 3, we illustrate the optimal parameters' settings of the ML models that we experimented with.

Table 3. Machine learning models' settings.

Models	Parameters	Models	Parameters
NB	useKernelEstimator: False useSupervisedDiscretization: True	**RotF**	classifier: RF numberOfGroups: True projectionFilter: PrincipalComponents
LR	ridge = 10^{-8} useConjugateGradientDescent: True	**J48**	reducedErrorPruning: False savelnstanceData: True useMDLCorrection : True, subtreeRaising: True binarySplits = True, collapseTree = True
MLP	learning rate = 0.1 momentum = 0.2 training time = 200	**Stacking**	classifiers: RF and NB metaClassifier: LR
KNN	K=3 Search Algorithm: LinearNNSearch with Euclidean cross-validate = True	**Voting**	classifiers: RF and NB combinationRule: average of probabilities
RF	breakTiesRadomly :True numIterations = 500 storeOutOfBagPredictions: True	**Bagging**	classifiers: RF printClassifiers : True storeOutOfBagPredictions: True

2.4. Evaluation Metrics

In order to evaluate the ML models' performance, we relied on the accuracy, precision, recall and AUC metrics [63]. The confusion matrix consists of the elements true positive (TP), true negative (TN), false positive (FP) and false-negative (FN). The aforementioned metrics are defined as follows:

- Accuracy: Summarizes the performance of the classification task and measures the number of correctly predicted instances out of all the data instances.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + +\text{FN} + \text{TP} + \text{FP}} \quad (2)$$

- Precision: Shows the ratio of positive subjects in relation to true and false positive subjects.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

- Recall: Corresponds to the proportion of participants who were diagnosed with CAD and were correctly considered positive, concerning all positive participants.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

- In order to evaluate the distinguishability of a model, the AUC is exploited. It is a metric that varies in [0, 1]. The closer to one, the better the ML model performance is in distinguishing CAD from non-CAD instances.

2.5. Experimental Setup

For the evaluation of our ML models, we relied on the Waikato environment for knowledge analysis (Weka) [64]. In addition, the experiments were performed on a computer system with the following specifications: 11th generation Intel(R) Core(TM) i7-1165 G7 @ 2.80GHz, RAM 16 GB, Windows 11 Home, 64-bit OS and x64 processor. We applied 10-fold cross-validation in order to measure the ML models' efficiency in the balanced dataset of 6198 instances after SMOTE, and in the unbalanced dataset of 3655 instances without SMOTE.

3. Results

The purpose of our evaluation is to highlight the role of the SMOTE technique in terms of developing ML models of high reliability and accuracy. In this direction, we experimented with well-known ML models, such as NB, LR, RotF, MLP, KNN, J48, bagging, RF, voting and stacking, evaluating them in terms of accuracy, recall, precision and AUC after 10-fold cross-validation with and without the use of SMOTE.

Specifically, the initial dataset includes 3655 instances. The number of participants who have been diagnosed with CAD is 556 (15.2%), while the non-CAD participants are 3099 (84.8%). According to Table 4 and without the application of SMOTE, the ML models we experimented with have quite high accuracy rates (as this metric captures the overall classification performance in both states of the class label) and less good rates in terms of AUC. AUC is a measure that shows the separation ability of a model among the distributions of CAD and non-CAD instances. The smaller their overlap is, the higher the AUC values will be. In the current dataset, the AUC values without SMOTE reveal that the models have a chance between 55.4% (KNN) and 71.3% (RotF) of being able to distinguish between CAD and Non-CAD classes. Moreover, focusing on the values of the Recall metric, which captures how many of the samples belonging to the CAD class were correctly classified, these ones are significantly low, ranging from 4.5% (bagging) to 31.8% (NB).

Furthermore, from Table 4, we see that after applying SMOTE, the ML models achieved very high-performance metrics. Focusing on the acquired experimental outcomes of the recall metric, its superiority over the No-SMOTE case is significant due to the reduction in

false-negative predictions. This is of great importance and plays a decisive role in the design of efficient ML models and techniques. The ratio of correctly recognized CAD samples ranges from 74.2% (LR) to 87.6% (stacking). The accuracy was less enhanced by the application of SMOTE, while the highest improvement of 10.6% is observed by the NB classifier.

Table 4. Performance evaluation of ML models in terms of accuracy, precision, recall and AUC metrics.

	Accuracy		Precision (CAD class)		Recall (CAD Class)		AUC	
	No SMOTE	**SMOTE**	**No SMOTE**	**SMOTE**	**No SMOTE**	**SMOTE**	**No SMOTE**	**SMOTE**
NB	0.700	0.906	0.336	0.973	0.318	0.835	0.700	0.941
LR	0.754	0.779	0.645	0.710	0.088	0.762	0.729	0.793
MLP	0.730	0.798	0.355	0.742	0.146	0.801	0.661	0.833
3-NN	0.722	0.796	0.311	0.760	0.140	0.867	0.585	0.854
RF	0.748	0.855	0.493	0.844	0.063	0.871	0.693	0.931
RotF	0.751	0.845	0.625	0.827	0.054	0.872	0.713	0.925
J48	0.714	0.787	0.268	0.777	0.205	0.804	0.636	0.857
Stacking	0.747	0.909	0.482	0.967	0.059	0.876	0.698	0.961
Bagging	0.748	0.843	0.500	0.827	0.045	0.866	0.702	0.926
Voting	0.787	0.908	0.367	0.960	0.187	0.852	0.702	0.958

To further interpret the classification performance of ML models, AUC–ROC curves are plotted in Figures 3 and 4, before and after the application of SMOTE. These are probability curves that capture the relationship between the true positive rate (TPR or recall) and the false positive rate (FPR), where FPR is defined as the ratio $\frac{FP}{FP+TN}$. As the results indicate, the SMOTE benefited most of the models by significantly improving the recall of the CAD class; thus, the AUC curves of ensemble models became more abrupt, starting from lower values of FPR and attaining one. As a final note, it is observed that after class balancing, stacking and voting have identical AUC curves, with a small lead in stacking in all metrics.

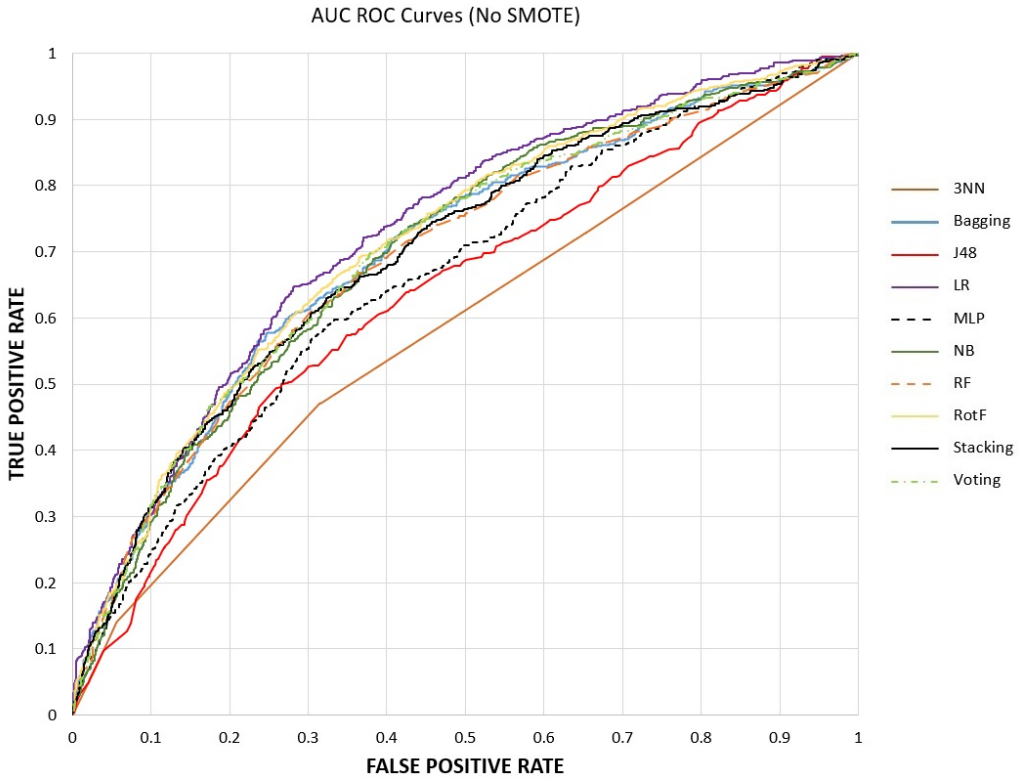

Figure 3. Performance Evaluation with AUC ROC Curves before SMOTE.

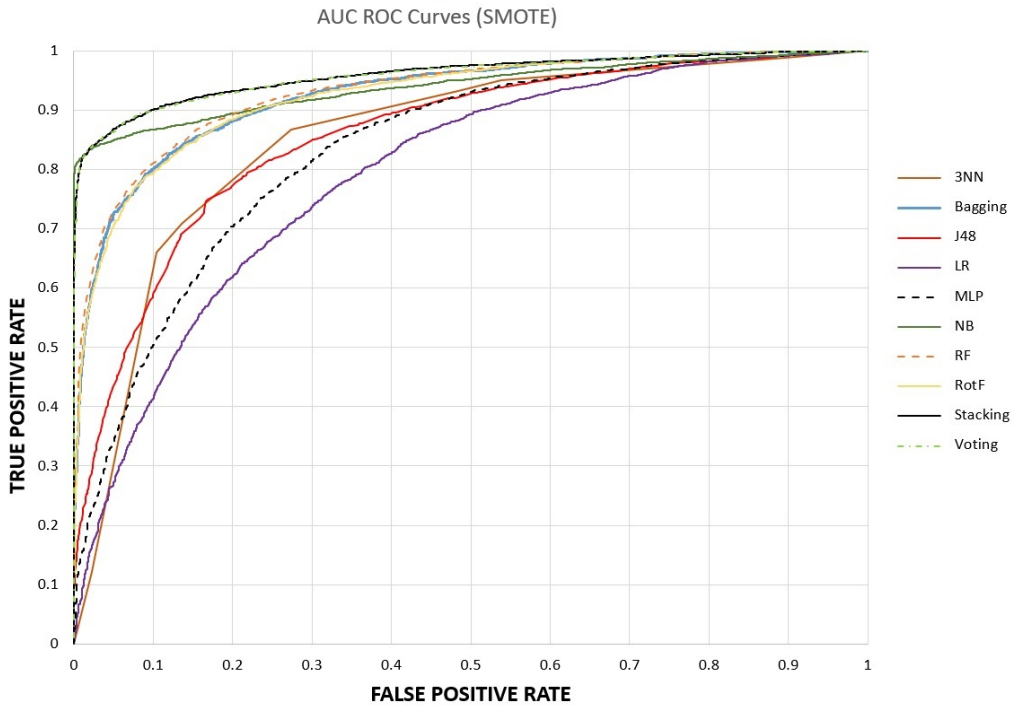

Figure 4. Performance Evaluation with AUC ROC Curves after SMOTE.

Concluding the results section, we should highlight that the stacking ensemble model after SMOTE with 10-fold cross-validation prevailed over the other models, achieving an accuracy of 90.9%, a precision of 96.7%, a recall of 87.6% and an AUC equal to 96.1%.

4. Discussion

In this section, a brief description of relevant works on coronary artery disease risk prediction is provided with the contribution of ML models and techniques.

First, the authors in [65] tested ten traditional ML algorithms. They also introduced a new optimization technique called the N2 Genetic optimizer. The experiments demonstrated that N2 Genetic-nuSVM provided an accuracy of 93.08% and an F1 score of 91.51% when predicting CAD outcomes among the patients included in the Z-Alizadeh Sani dataset.

Similarly, the authors in [66] used the publicly available Z-Alizadeh Sani dataset, which contains random samples of 216 cases with CAD and 87 normal controls with 56 different features. Five different supervised classification ML algorithms, LR, a classification tree with bagging (bagging CART), RF, SVM, and KNN, were applied. Finally, the results indicate that the SVM model is able to predict the presence of CAD more effectively and accurately than other models, with an accuracy of 89.4%, a sensitivity of 94.3%, a specificity of 78.2%, and an AUC of 88.7%.

In addition, the authors in [67] compare the accurate prediction results of NB and SVM in order to predict CAD in a timely manner. This research paper uses two types of datasets (noisy and less noisy images along with numerical features), where the models experimented with them. The NB model has lower accuracy compared to the SVM in both cases.

In [68], the authors applied ML algorithms, including SVM, KNN, RT, RF, NB, gradient boosting (GB) and LR, on a dataset obtained in the two General Hospitals in Kano State, Nigeria for the prediction of CAD. In terms of accuracy, the random forest model emerged as the best model with 92.04%; for specificity, the NB model was the best, with 92.40%. For sensitivity, the SVM model was the best, with 87.34%, and for the AUC, the best model was the RF model, with 92.20%.

The research study in [69] aimed to improve the accuracy of CAD diagnosis by selecting the most significant features. For this purpose, several ML models such as the RT, the C5.0 DT and the SVM, were evaluated. The RT showed promising results, achieving the highest accuracy of 91.47%.

Moreover, in [70], the LR, SVM and ANN algorithms are the points of interest. In order to evaluate the results, the accuracy and AUC scores have been performed using the 10-fold cross-validation. The SMOTE technique has been used to balance the dataset. The ANN achieved the highest accuracy of 93.35% and an AUC of 98% for CAD prediction.

Furthermore, the same methodology is followed by the authors in [71]. Three feature selection methods have been used on 13 input features from the Cleveland dataset with 297 entries, and 7 were selected. Specifically, SVM, NB and KNN using 10-fold cross-validation were applied for CAD prediction. The NB classifier performs the best on this dataset, achieving an accuracy of 84%.

Furthermore, the authors in [72–74] experimented with the same dataset [34] as the current study. In [72], the neural network is the algorithm that yielded the greatest AUC in R-studio when excluding observations in which there was at least one missing value (AUC = 71%). When the data was analyzed in RapidMiner, the best algorithm was SVM (AUC = 75%).

The study in [73] applied LR, NB, DT, KNN, SVM and RF in order to predict whether a subject runs a risk of future development of CAD or not in the next ten years. The RF model outperformed the other models with an accuracy of 91.1%, a precision of 64.3% and a recall equal to 6.4%. In [74], work suggested the cloud RF (C-RF) model, which prevailed compared to CART (classification and regression tree), SVM and CNN, with accuracy and an AUC of 85%, similarly.

Here, in the balanced dataset after SMOTE, we exploited more efficient schemes to design the desired classification models, with an emphasis on ensemble techniques. Furthermore, we further validated the expected performance of ensemble models with a graphical illustration of the AUC–ROC curves. To sum up, comparing the performance of [72–74], our proposed trained and tested classifier (i.e., stacking) presents an accuracy of 90.9 %, a precision of 96.7%, a recall of 87.6% and an AUC equal to 96.1% after SMOTE with 10-fold cross-validation, thus confirming its high accuracy rates.

5. Conclusions

Cardiovascular disease remains the leading cause of death despite significant progress in medical science and contains a wide range of diseases, including all pathological changes involving the heart and/or blood vessels. The long-term risk prediction of CAD disease was the topic under consideration in this research. Furthermore, the features' importance evaluation, based on the gain ratio and RF, was performed. Through risk factor monitoring and analysis, personalized guidelines and interventions can be suggested to prevent CAD occurrence. Such an analysis can help medical experts regularly reassess underlying risks, and even if CAD occurs, they can provide patients with novel guidelines and treatments based on individual patient characteristics, that may enhance their daily life, increase life expectancy and restrict mortality.

Furthermore, experimental evaluation with various ML models, including NB, LR, RotF, MLP, KNN, J48, bagging, RF, voting and stacking with 10-fold cross-validation, after the use or not of SMOTE, was made. Comparing the ML models in terms of accuracy, recall, precision and AUC, the reliability of the SMOTE technique was demonstrated. The stacking ensemble model after SMOTE with 10-fold cross-validation was the model that prevailed over the other ones, achieving an accuracy of 90.9%, a precision of 96.7%, a recall of 87.6%, and an AUC equal to 96.1%; this constitutes the main proposition of this paper. In future work, we aim to extend the machine learning framework by using deep learning methods and comparing the results based on the aforementioned metrics.

Author Contributions: E.D. and M.T. conceived of the idea, designed and performed the experiments, analyzed the results, drafted the initial manuscript and revised the final manuscript. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Buijtendijk, M.F.; Barnett, P.; van den Hoff, M.J. Development of the human heart. *Am. J. Med. Genet. Part C Semin. Med. Genet.* **2020**, *184*, 7–22. [CrossRef] [PubMed]
2. Lopez, E.O.; Ballard, B.D.; Jan, A. Cardiovascular disease. In *StatPearls [Internet]*; StatPearls Publishing: Tampa, FL, USA, 2021.
3. Pagliaro, B.R.; Cannata, F.; Stefanini, G.G.; Bolognese, L. Myocardial ischemia and coronary disease in heart failure. *Heart Fail. Rev.* **2020**, *25*, 53–65. [CrossRef] [PubMed]
4. Malakar, A.K.; Choudhury, D.; Halder, B.; Paul, P.; Uddin, A.; Chakraborty, S. A review on coronary artery disease, its risk factors, and therapeutics. *J. Cell. Physiol.* **2019**, *234*, 16812–16823. [CrossRef] [PubMed]
5. Fox, K.A.; Metra, M.; Morais, J.; Atar, D. The myth of 'stable'coronary artery disease. *Nat. Rev. Cardiol.* **2020**, *17*, 9–21. [CrossRef]
6. Lee, S.E.; Sung, J.M.; Rizvi, A.; Lin, F.Y.; Kumar, A.; Hadamitzky, M.; Kim, Y.J.; Conte, E.; Andreini, D.; Pontone, G.; et al. Quantification of coronary atherosclerosis in the assessment of coronary artery disease. *Circ. Cardiovasc. Imaging* **2018**, *11*, e007562. [CrossRef]
7. Reeh, J.; Therming, C.B.; Heitmann, M.; Højberg, S.; Sørum, C.; Bech, J.; Husum, D.; Dominguez, H.; Sehestedt, T.; Hermann, T.; et al. Prediction of obstructive coronary artery disease and prognosis in patients with suspected stable angina. *Eur. Heart J.* **2019**, *40*, 1426–1435. [CrossRef]
8. Goyal, A.; Zeltser, R. Unstable angina. In *StatPearls [Internet]*; StatPearls Publishing: Tampa, FL, USA, 2022.
9. Shao, C.; Wang, J.; Tian, J.; Tang, Y.d. Coronary artery disease: From mechanism to clinical practice. *Adv. Exp. Med. Biol.* **2020**, *1177*, 1–36.
10. Wong, C.X.; Brown, A.; Lau, D.H.; Chugh, S.S.; Albert, C.M.; Kalman, J.M.; Sanders, P. Epidemiology of sudden cardiac death: Global and regional perspectives. *Heart Lung Circ.* **2019**, *28*, 6–14. [CrossRef]
11. Nowbar, A.N.; Gitto, M.; Howard, J.P.; Francis, D.P.; Al-Lamee, R. Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. *Circ. Cardiovasc. Qual. Outcomes* **2019**, *12*, e005375. [CrossRef]
12. Mensah, G.A.; Roth, G.A.; Fuster, V. The global burden of cardiovascular diseases and risk factors: 2020 and beyond. *J. Am. Coll. Cardiol.* **2019**, *74*, 2529–2532. [CrossRef]
13. Ambrose, J.A.; Najafi, A. Strategies for the prevention of coronary artery disease complications: Can we do better? *Am. J. Med.* **2018**, *131*, 1003–1009. [CrossRef] [PubMed]
14. Houston, M. The role of noninvasive cardiovascular testing, applied clinical nutrition and nutritional supplements in the prevention and treatment of coronary heart disease. *Ther. Adv. Cardiovasc. Dis.* **2018**, *12*, 85–108. [CrossRef]
15. Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access* **2021**, *9*, 103737–103757. [CrossRef]
16. Dritsas, E.; Trigka, M. Data-driven machine-learning methods for diabetes risk prediction. *Sensors* **2022**, *22*, 5304. [CrossRef] [PubMed]
17. Alexiou, S.; Dritsas, E.; Kocsis, O.; Moustakas, K.; Fakotakis, N. An approach for Personalized Continuous Glucose Prediction with Regression Trees. In Proceedings of the 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Preveza, Greece, 24–26 September 2021; pp. 1–6.
18. Dritsas, E.; Alexiou, S.; Konstantoulas, I.; Moustakas, K. Short-term Glucose Prediction based on Oral Glucose Tolerance Test Values. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022)—Volume 5: HEALTHINF, Online Streaming, 9–11 February 2022; pp. 249–255.
19. Fazakis, N.; Dritsas, E.; Kocsis, O.; Fakotakis, N.; Moustakas, K. Long-Term Cholesterol Risk Prediction with Machine Learning Techniques in ELSA Database. In Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI), SCIPTRESS, Online Streaming, 25–27 October 2021; pp. 445–450.
20. Dritsas, E.; Trigka, M. Machine learning methods for hypercholesterolemia long-term risk prediction. *Sensors* **2022**, *22*, 5365. [CrossRef] [PubMed]
21. Dritsas, E.; Fazakis, N.; Kocsis, O.; Fakotakis, N.; Moustakas, K. Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database. In Proceedings of the International Conference on Learning and Intelligent Optimization, Athens, Greece, 20–25 June 2021; pp. 113–120.
22. Dritsas, E.; Alexiou, S.; Moustakas, K. Efficient Data-driven Machine Learning Models for Hypertension Risk Prediction. In Proceedings of the 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Biarritz, France, 8–12 August 2022; pp. 1–6.

23. Dritsas, E.; Alexiou, S.; Moustakas, K. COPD severity prediction in elderly with ML techniques. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu Greece, 29 June–1 July 2022; pp. 185–189.
24. Dritsas, E.; Trigka, M. Supervised Machine Learning Models to Identify Early-Stage Symptoms of SARS-CoV-2. *Sensors* **2022**, *23*, 40. [CrossRef]
25. Dritsas, E.; Trigka, M. Stroke risk prediction with machine learning techniques. *Sensors* **2022**, *22*, 4670. [CrossRef]
26. Dritsas, E.; Trigka, M. Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cogn. Comput.* **2022**, *6*, 98. [CrossRef]
27. Dritsas, E.; Trigka, M. Supervised Machine Learning Models for Liver Disease Risk Prediction. *Computers* **2023**, *12*, 19. [CrossRef]
28. Konstantoulas, I.; Kocsis, O.; Dritsas, E.; Fakotakis, N.; Moustakas, K. Sleep Quality Monitoring with Human Assisted Corrections. In Proceedings of the International Joint Conference on Computational Intelligence (IJCCI), SCIPTRESS, Online Streaming, 25–27 October 2021; pp. 435–444.
29. Konstantoulas, I.; Dritsas, E.; Moustakas, K. Sleep Quality Evaluation in Rich Information Data. In Proceedings of the 2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA), Corfu, Greece, 18–20 July 2022; pp. 1–4.
30. Konerman, M.A.; Beste, L.A.; Van, T.; Liu, B.; Zhang, X.; Zhu, J.; Saini, S.D.; Su, G.L.; Nallamothu, B.K.; Ioannou, G.N.; et al. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PLoS ONE* **2019**, *14*, e0208141. [CrossRef]
31. Dritsas, E.; Alexiou, S.; Moustakas, K. Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. In Proceedings of the ICT4AWE 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health, Online Streaming, 22–24 April 2022; pp. 315–321.
32. Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. *Big Data Cogn. Comput.* **2022**, *6*, 139. [CrossRef]
33. Dritsas, E.; Alexiou, S.; Moustakas, K. Metabolic Syndrome Risk Forecasting on Elderly with ML Techniques. In Proceedings of the 16th International Conference on Learning and Intelligent Optimization, Cyclades, Greece, 5–10 June 2022.
34. Coronary Prediction Dataset. Available online: https://www.kaggle.com/datasets/jiantay33/coronary-prediction (accessed on 27 December 2022).
35. Wada, H.; Miyauchi, K.; Daida, H. Gender differences in the clinical features and outcomes of patients with coronary artery disease. *Expert Rev. Cardiovasc. Ther.* **2019**, *17*, 127–133. [CrossRef] [PubMed]
36. Tillmann, T.; Vaucher, J.; Okbay, A.; Pikhart, H.; Peasey, A.; Kubinova, R.; Pajak, A.; Tamosiunas, A.; Malyutina, S.; Hartwig, F.P.; et al. Education and coronary heart disease: Mendelian randomisation study. *BMJ* **2017**, *358*. [CrossRef] [PubMed]
37. Kondo, T.; Nakano, Y.; Adachi, S.; Murohara, T. Effects of tobacco smoking on cardiovascular disease. *Circ. J.* **2019**, *83*, 1980–1985. [CrossRef]
38. Fuchs, F.D.; Whelton, P.K. High blood pressure and cardiovascular disease. *Hypertension* **2020**, *75*, 285–292. [CrossRef]
39. Katsanos, A.H.; Palaiodimou, L.; Price, C.; Giannopoulos, S.; Lemmens, R.; Kosmidou, M.; Georgakis, M.; Weimar, C.; Kelly, P.; Tsivgoulis, G. Colchicine for stroke prevention in patients with coronary artery disease: A systematic review and meta-analysis. *Eur. J. Neurol.* **2020**, *27*, 1035–1038. [CrossRef]
40. Vidal-Petiot, E.; Greenlaw, N.; Ford, I.; Ferrari, R.; Fox, K.M.; Tardif, J.C.; Tendera, M.; Parkhomenko, A.; Bhatt, D.L.; Steg, P.G.; et al. Relationships between components of blood pressure and cardiovascular events in patients with stable coronary artery disease and hypertension. *Hypertension* **2018**, *71*, 168–176. [CrossRef]
41. Fishman, S.L.; Sonmez, H.; Basman, C.; Singh, V.; Poretsky, L. The role of advanced glycation end-products in the development of coronary artery disease in patients with and without diabetes mellitus: A review. *Mol. Med.* **2018**, *24*, 59. [CrossRef] [PubMed]
42. Hackshaw, A.; Morris, J.K.; Boniface, S.; Tang, J.L.; Milenković, D. Low cigarette consumption and risk of coronary heart disease and stroke: Meta-analysis of 141 cohort studies in 55 study reports. *BMJ* **2018**, *361*, k1611. [CrossRef]
43. Tada, H.; Nohara, A.; Inazu, A.; Sakuma, N.; Mabuchi, H.; Kawashiri, M.A. Sitosterolemia, hypercholesterolemia, and coronary artery disease. *J. Atheroscler. Thromb.* **2018**, *25*, 783–789. [CrossRef]
44. Nazarzadeh, M.; Pinho-Gomes, A.C.; Byrne, K.S.; Canoy, D.; Raimondi, F.; Solares, J.R.A.; Otto, C.M.; Rahimi, K. Systolic blood pressure and risk of valvular heart disease: A Mendelian randomization study. *JAMA Cardiol.* **2019**, *4*, 788–795. [CrossRef] [PubMed]
45. Tackling, G.; Borhade, M.B. Hypertensive heart disease. In *StatPearls [Internet]*; StatPearls Publishing: Tampa, FL, USA, 2021.
46. Piché, M.E.; Tchernof, A.; Després, J.P. Obesity phenotypes, diabetes, and cardiovascular diseases. *Circ. Res.* **2020**, *126*, 1477–1500. [CrossRef] [PubMed]
47. Forte, G.; Favieri, F.; Casagrande, M. Heart rate variability and cognitive function: A systematic review. *Front. Neurosci.* **2019**, *13*, 710. [CrossRef]
48. Xia, J.; Yin, C. Glucose variability and coronary artery disease. *Heart Lung Circ.* **2019**, *28*, 553–559. [CrossRef] [PubMed]
49. Rattan, V.; Mittal, R.; Singh, J.; Malik, V. Analyzing the Application of SMOTE on Machine Learning Classifiers. In Proceedings of the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India , 5–7 March 2021; pp. 692–695.
50. Cunningham, P.; Delany, S.J. k-Nearest neighbour classifiers-A Tutorial. *Acm Comput. Surv. (CSUR)* **2021**, *54*, 1–25. [CrossRef]

51. Dritsas, E.; Fazakis, N.; Kocsis, O.; Moustakas, K.; Fakotakis, N. Optimal Team Pairing of Elder Office Employees with Machine Learning on Synthetic Data. In Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 12–14 July 2021; pp. 1–4.
52. Gnanambal, S.; Thangaraj, M.; Meenatchi, V.; Gayathri, V. Classification algorithms with attribute selection: An evaluation study using WEKA. *Int. J. Adv. Netw. Appl.* **2018**, *9*, 3640–3644.
53. Dimitriadis, S.I.; Liparas, D.; Tsolaki, M.N.; Initiative, A.D.N.; et al. Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) database. *J. Neurosci. Methods* **2018**, *302*, 14–23.
54. Berrar, D. Bayes' theorem and naive Bayes classifier. *Encycl. Bioinform. Comput. Biol. ABC Bioinform.* **2018**, *403* .
55. Nusinovici, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.Y. Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **2020**, *122*, 56–69. [CrossRef]
56. Posonia, A.M.; Vigneshwari, S.; Rani, D.J. Machine Learning based Diabetes Prediction using Decision Tree J48. In Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 3–5 December 2020; pp. 498–502.
57. González, S.; García, S.; Del Ser, J.; Rokach, L.; Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **2020**, *64*, 205–237. [CrossRef]
58. Palimkar, P.; Shaw, R.N.; Ghosh, A. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced Computing and Intelligent Technologies*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 219–244.
59. Husna, N.A.; Bustamam, A.; Yanuar, A.; Sarwinda, D. The drug design for diabetes mellitus type II using rotation forest ensemble classifier. *Procedia Comput. Sci.* **2021**, *179*, 161–168. [CrossRef]
60. Dogan, A.; Birant, D. A weighted majority voting ensemble approach for classification. In Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 11–15 September 2019; pp. 1–6.
61. Pavlyshenko, B. Using stacking approaches for machine learning models. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 255–258.
62. Masih, N.; Naz, H.; Ahuja, S. Multilayer perceptron based deep neural network for early detection of coronary heart disease. *Health Technol.* **2021**, *11*, 127–138. [CrossRef]
63. Handelman, G.S.; Kok, H.K.; Chandra, R.V.; Razavi, A.H.; Huang, S.; Brooks, M.; Lee, M.J.; Asadi, H. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *Am. J. Roentgenol.* **2019**, *212*, 38–43. [CrossRef] [PubMed]
64. Weka. Available online: https://www.weka.io/ (accessed on 27 December 2022).
65. Abdar, M.; Książek, W.; Acharya, U.R.; Tan, R.S.; Makarenkov, V.; Pławiak, P. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* **2019**, *179*, 104992. [CrossRef] [PubMed]
66. Dahal, K.R.; Gautam, Y. Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. *Open J. Stat.* **2020**, *10*, 694–705. [CrossRef]
67. Chen, J.I.Z.; Hengjinda, P. Early prediction of coronary artery disease (cad) by machine learning method-a comparative study. *J. Artif. Intell.* **2021**, *3*, 17–33.
68. Muhammad, L.; Al-Shourbaji, I.; Haruna, A.A.; Mohammed, I.; Ahmad, A.; Jibrin, M.B. Machine Learning Predictive Models for Coronary Artery Disease. *Sn Comput. Sci.* **2021**, *2*, 350. [CrossRef]
69. Joloudari, J.H.; Hassannataj Joloudari, E.; Saadatfar, H.; Ghasemigol, M.; Razavi, S.M.; Mosavi, A.; Nabipour, N.; Shamshirband, S.; Nadai, L. Coronary artery disease diagnosis; ranking the significant features using a random trees model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 731. [CrossRef]
70. Dipto, I.C.; Islam, T.; Rahman, H.M.; Rahman, M.A. Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. *J. Data Anal. Inf. Process.* **2020**, *8*, 41–68. [CrossRef]
71. Nassif, A.B.; Mahdi, O.; Nasir, Q.; Talib, M.A.; Azzeh, M. Machine learning classifications of coronary artery disease. In Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 15–17 November 2018, pp. 1–6.
72. Beunza, J.J.; Puertas, E.; García-Ovejero, E.; Villalba, G.; Condes, E.; Koleva, G.; Hurtado, C.; Landecho, M.F. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *J. Biomed. Inform.* **2019**, *97*, 103257. [CrossRef]
73. Minou, J.; Mantas, J.; Malamateniou, F.; Kaitelidou, D. Classification Techniques for Cardio-Vascular Diseases Using Supervised Machine Learning. *Med. Arch.* **2020**, *74*, 39. [CrossRef] [PubMed]
74. Wang, J.; Rao, C.; Goh, M.; Xiao, X. Risk assessment of coronary heart disease based on cloud-random forest. *Artif. Intell. Rev.* **2023**, *56*, 203–232. [CrossRef]