electronics

MDPI

Article

Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology

Shasha Zhang [1], Yuyu Yuan [1,2], Zhonghua Yao [3], Xinyan Wang [4] and Zhen Lei [4,*]

[1] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China; Shashazhang@bupt.edu.cn (S.Z.); yuanyuyu@bupt.edu.cn (Y.Y.)
[2] Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing 100876, China
[3] School of Computer and Engineering, Fuyang Normal University, Fuyang 236000, China; yaozh@fynu.edu.cn
[4] Air Force Medical Center, PLA, Beijing 100142, China; wxy61092@163.com
* Correspondence: leizhen61019@163.com

Abstract: Coronary artery disease (CAD) is one of the diseases with the highest morbidity and mortality in the world. In 2019, the number of deaths caused by CAD reached 9.14 million. The detection and treatment of CAD in the early stage is crucial to save lives and improve prognosis. Therefore, the purpose of this research is to develop a machine-learning system that can be used to help diagnose CAD accurately in the early stage. In this paper, two classical ensemble learning algorithms, namely, XGBoost algorithm and Random Forest algorithm, were used as the classification model. In order to improve the classification accuracy and performance of the model, we applied four feature processing techniques to process features respectively. In addition, synthetic minority oversampling technology (SMOTE) and adaptive synthetic (ADASYN) were used to balance the dataset, which included 71.29% CAD samples and 28.71% normal samples. The four feature processing technologies improved the performance of the classification models in terms of classification accuracy, precision, recall, F_1 score and specificity. In particular, the XGBboost algorithm achieved the best prediction performance results on the dataset processed by feature construction and the SMOTE method. The best classification accuracy, recall, specificity, precision, F_1 score and AUC were 94.7%, 96.1%, 93.2%, 93.4%, 94.6% and 98.0%, respectively. The experimental results prove that the proposed method can accurately and reliably identify CAD patients from suspicious patients in the early stage and can be used by medical staff for auxiliary diagnosis.

Keywords: cardiovascular disease; coronary artery disease; feature smoothing; feature encoding; feature selection; feature construction; XGBoost; classification

Citation: Zhang, S.; Yuan, Y.; Yao, Z.; Wang, X.; Lei, Z. Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology. *Electronics* **2022**, *11*, 315. https://doi.org/10.3390/electronics11030315

Academic Editor: Jun-Ho Huh

Received: 13 December 2021
Accepted: 17 January 2022
Published: 20 January 2022

1. Introduction

Cardiovascular diseases (CVDs) are the main causes of death in the world. In 2019, the number of deaths caused by cardiovascular diseases reached 18.5 million, accounting for about one third of the total deaths in the world [1,2]. Among them, nearly half of the deaths caused by cardiovascular diseases are caused by coronary artery disease (CAD). Coronary artery disease (CAD) is regarded as one of the most usual types of cardiovascular diseases. In 2019, there were 197 million CAD patients worldwide [1,3].

CAD refers to the stenosis or occlusion of the coronary arteries due to atherosclerotic changes, which prevents the oxygen-rich blood flow from entering the heart, leading to ischemic heart attacks. According to the anatomy of the coronary arteries, there are three main blood vessels supplying blood to the myocardium, namely, (1) the left anterior descending artery (LAD), (2) the left circumflex artery (LCX), and (3) the right coronary artery (RCA). CAD occurs when any one of the blood vessels is blocked by more than

50% [4]. A large number of medical studies have confirmed that early detection and treatment of CAD is essential to save lives and improve prognosis. Therefore, many doctors and scholars are committed to finding methods that can produce the early detection and diagnosis of CAD; machine learning and data mining technologies are one of them. In recent years, machine learning and data mining technologies that construct predictive models by extracting hidden relationships between features and diseases from available clinical datasets have been widely used for disease screening, risk stratification, prediction, and decision-making assistance [5–7], and usually achieve better predictive performance.

The performance of the model constructed by machine learning and data mining technologies is often determined by features and algorithms. Many feature processing technologies such as feature selection and feature construction have been applied in the existing literature. Generally speaking, the selection of relevant feature subsets has the potential to improve model accuracy and test performance. Therefore, many studies have tried different feature selection methods. The most frequently used methods are information gain, weight by SVM, PCA and Gini coefficient [8]. The applications of other feature selection algorithms are as follows: Elham Nasarian et al. developed an algorithm called heterogeneous hybrid feature selection (2HFs) for CAD detection. On the Nasarian CAD dataset, the proposed feature selection method achieved 81.23% classification accuracy with XGBoost classifier [9]. Moloud Abdar et al. used the genetic algorithm and particle swarm optimization algorithm to eliminate redundant features, which significantly enhanced the performance of traditional algorithms [10]. Burak Kolukisa et al. used four feature selection methods: information gain, gain ratio, Relief-F and chi-squared, combined with linear discriminant analysis, to process three public heart disease datasets, and the prediction results were better than the feature selection method based on doctor experience [11]. Mariam Zomorodi-Moghadam et al. built a rule set based on a multi-objective evolutionary search and particle swarm optimization (PSO) algorithm to predict CAD, and achieved 90% classification accuracy [12]. Zeinab Arabasadi et al. improved the initial weight value of the neural network through a genetic algorithm, and improved the classification performance of the neural network by about 10% [13]. Joloudari et al. proposed a method that integrates SVM, random trees (RTs), C5.0 and chi-squared automatic interaction detection (CHAID), four machine learning algorithms used for selecting the features most relevant to CAD prediction and classifying the instances. Finally, this method obtained the best results by using the RTs model, and the classification accuracy was 91.47% [14]. Data preprocessing and feature construction are also critical technologies to discover the potential laws of data and improve the prediction ability of models. The most commonly used data preprocessing methods are discretization [15] and normalization [10]. Compared with feature selection, there is less application of feature construction in the existing literature. Excepting that [15,16] constructed three new features based on the stenosis of LAD, LCX and RCA vessels for CAD prediction, almost no other research has used the idea of feature construction. At the level of algorithm selection, the most widely used computational algorithms for CAD prediction are artificial neural network, decision tree, SVM, naive Bayes, KNN and system methods based on fuzzy rules [8]. In addition, ensemble learning technology is also used for CAD prediction. The authors of [17] used four ensemble learning techniques to boost the function of the base classifier on the Z-Alizadeh Sani and Cleveland datasets, and achieved 94.66% and 98.60% classification accuracy on the two datasets, respectively. Ashish et al. used SVM and XG-Boost algorithms for the identification of ischemic heart disease. The method produced classification accuracy of 93.8% and F_1 score of 91.8% [18].

In this paper, some new feature processing technologies such as feature smoothing, feature encoding, feature construction and feature selection are applied to the Z-Alizadeh Sani dataset to explore the method of detecting CAD quickly and accurately. Two classical ensemble learning algorithms, namely, XGBoost algorithm and Random Forest algorithm, are used as the classification model. Simultaneously, synthetic minority oversampling technology (SMOTE) and adaptive synthetic (ADASYN) are used to balance the dataset, which includes 71.29% CAD samples and 28.71% normal samples. In addition, 10-fold

cross-validation technology is used for testing the stability and accuracy of the model. Accuracy, recall, specificity, precision, F_1 score and AUC model evaluation measurements are applied to assess the power of the proposed model. The schematic diagram of the proposed method for CAD prediction is shown in Figure 1.

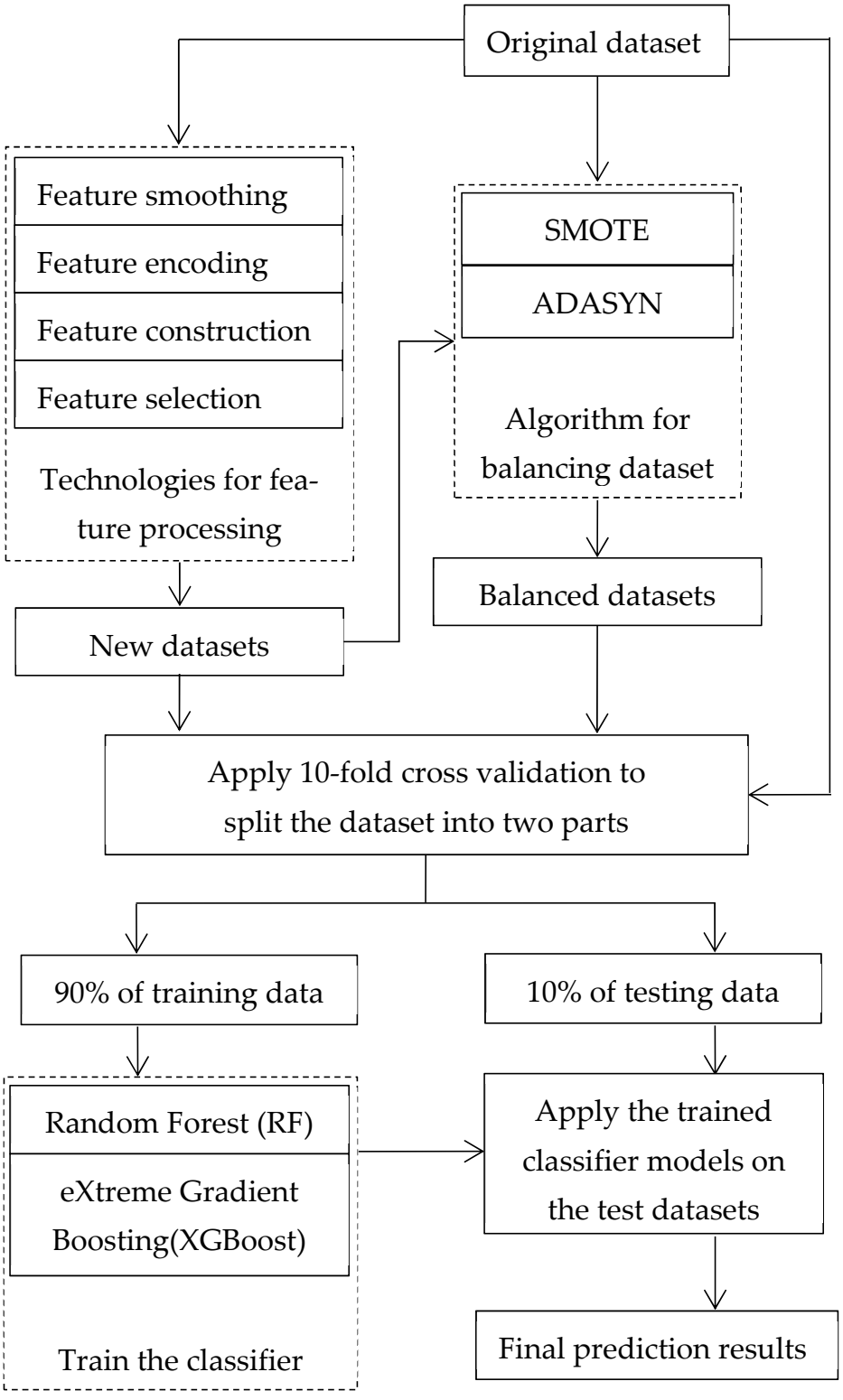

Figure 1. The schematic diagram for prediction of CAD.

The structure of the rest part of this paper is arranged as follows. An introduction to the proposed method is shown in Section 2. Section 3 provides details about the experiments and results. The classification performance of the proposed methods and our future works are discussed in Section 4. Finally, the conclusion is stated in Section 5.

2. Proposed Method

2.1. Feature Processing Technologies

2.1.1. Feature Smoothing

Feature smoothing is used to smooth the abnormal data contained in the input feature to a specific range. Since feature smoothing does not filter out or delete any records, the numbers of input features and samples remain unchanged after feature smoothing. Feature smoothing is divided into ZScore smoothing, percentile smoothing and threshold smoothing. This article uses Zscore smoothing to process the continuous features in the dataset used in this study. If the feature distribution follows a normal distribution, the noise is generally concentrated on the outside of $\mu \pm 3 * \sigma$. The visual expression of Zscore smoothing handling outliers in features is shown in the following formula:

$$x_{noisej} = \begin{cases} \mu + 3 * \sigma \; if \; x_{noisej} > \mu + 3 * \sigma \\ \mu - 3 * \sigma \; if \; x_{noisej} < \mu - 3 * \sigma \end{cases} \quad (1)$$

where x_{noisej} is the noisy data of feature j, μ is the mean of the feature j, and σ is the standard deviation of the feature j.

2.1.2. Feature Encoding

Feature frequency encoding is used to calculate the frequency of feature values appearance, and this frequency is used to replace feature values. Feature frequency encoding can express the probability information of feature appearance without changing the dimension of the dataset and losing feature information. At the same time, feature frequency encoding can avoid higher feature value dominance models. The original dataset contains both continuous features and categorical features. The value ranges of different continuous features are quite different. Categorical features include two-category and multi-category features. These characteristics of the dataset will affect the stability and convergence speed of the models. In addition, the dataset used in this paper also has the characteristics of a small sample size and many categorical features. For categorical features, the most commonly used encoding method is one-hot encoding. However, for this dataset, one-hot encoding will significantly increase the dimension of the dataset, and may lead to high-dimensional parallelism and multicollinearity. Therefore, this paper adopts feature frequency encoding to deal with continuous features and multi-category features in the dataset used in this study. The intuitive expression of feature frequency encoding is shown in Figure 2, where n_1, n_2, n_3 ... n_m are the frequencies corresponding to the values (x_{1j}, x_{2j}, x_{3j} ... x_{nj}) of feature j in the original dataset, respectively. Taking x_{1j} as an example, the values x_{1j} of feature j in the original dataset are replaced by the corresponding frequency n_1, which is used as a new feature value for training and testing.

2.1.3. Feature Construction

Feature construction refers to the artificial formation of some valuable features for prediction from the original data. We calculate the sum, mean and standard deviation of continuous features to form new features. Considering the influence of categorical features such as sex and age on continuous features, we calculate the sum, mean and standard deviation of continuous features based on different values of categorical features. For example, we generate triglyceride (TG) features based on sex feature; that is, the sum, mean and standard deviation of TG are calculated separately in male and female groups to form new features. The same continuous features calculation process is also carried out after the bucket division operation for age. Finally, the calculated new features are added to the raw

dataset to form a new dataset. The intuitive expression of feature construction is shown in Figure 3. Two columns named feature *j* and Sex on the left side of the figure represent two features in the original dataset. X_{nj} is the value of the nth sample on feature *j*. In the middle part of the figure, the values of feature *j* are grouped according to Sex. Where male = 0 means that the sample with male sex is coded with 0, and *m* is the number of males in the sample. Similarly, female = 1 means that the sample with female sex is coded with 1, and $n - m$ is the number of females in the sample. The three columns on the right side of the figure with the names of 0_j_sum, 0_j_mean and 0_j_std are the sum, mean and standard deviation of the values of feature *j* in male group. The corresponding values (e.g., $x_{male,j,sum}$, $x_{male,j,mean}$, $x_{male,j,std}$) of the three columns are new features' values after feature construction. The operation of the female group is the same as that of the male group. Finally, the new feature columns are obtained by merging the column with same names in the male group and the female group.

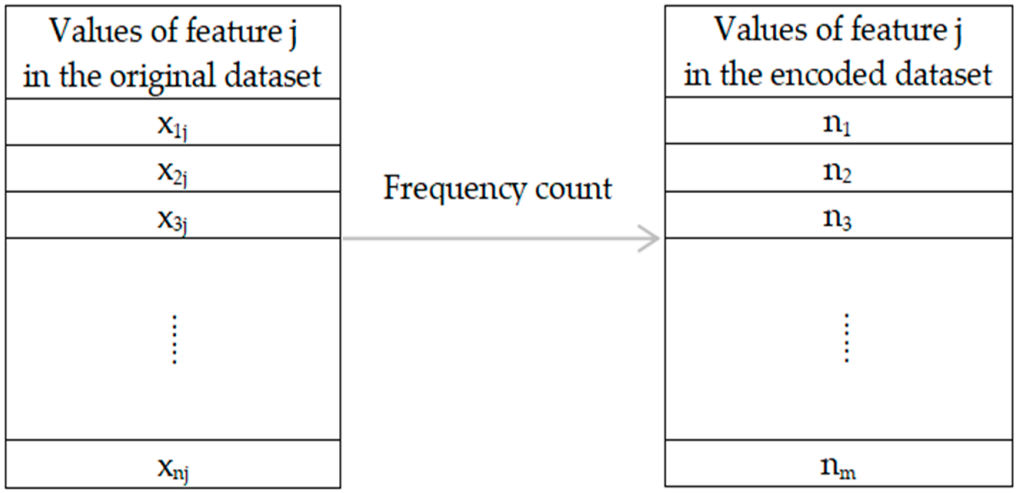

Figure 2. The schematic diagram of feature frequency encoding.

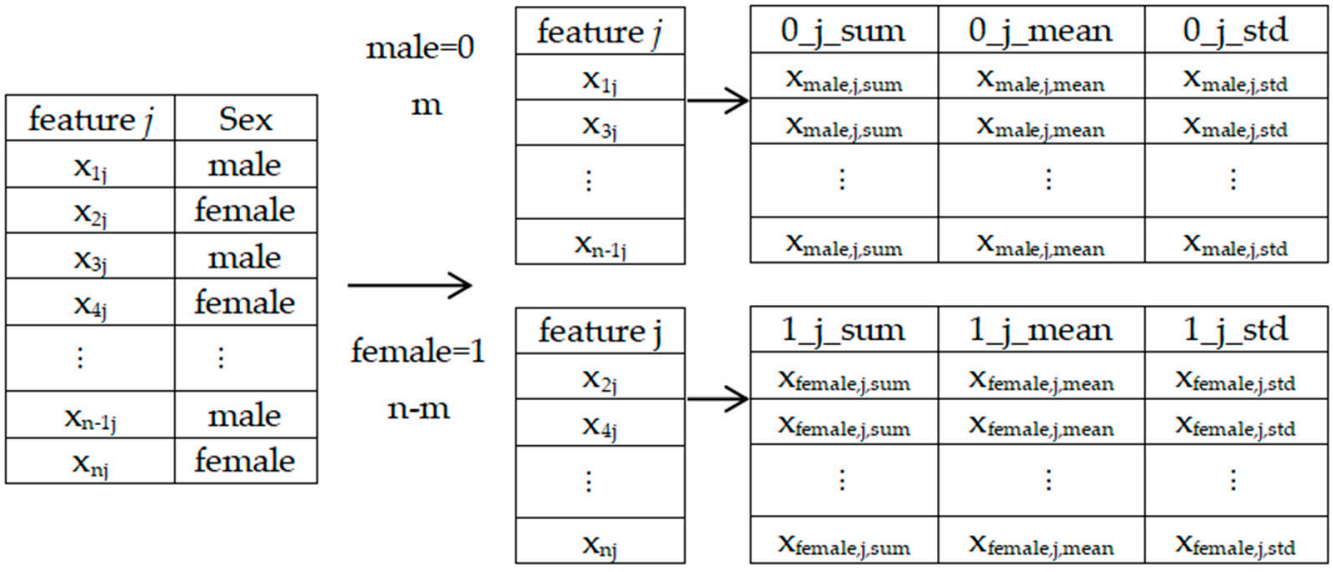

Figure 3. The schematic diagram of feature construction.

2.1.4. Feature Selection

Feature selection, a subset of feature engineering, acts as a pivotal part in enhancing the capacity of machine learning and data mining algorithms [15]. The major goal of feature selection is to select better features for prediction from the raw data. That is, for n features

in $x^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)}\right\}$, k $(k < n)$ features are selected from them to enhance the capacity of the machine learning algorithm. Through feature selection, the most relevant, important and less redundant feature subsets will be identified. Feature selection is the most widely used feature processing technology on the Z-Alizadeh Sani dataset in the existing literature. The most frequently used feature selection methods are information gain, weight by SVM, PCA and Gini coefficient. In our study, GBDT (Gradient Boosting Decision Tree) algorithm, which has been used as a feature selection in many prediction tasks and has achieved good results, is used for feature selection [19,20]. GBDT is an iterative decision tree algorithm. The algorithm is composed of multiple decision trees. By accumulating the prediction results of each decision tree, the final prediction conclusion of the algorithm is obtained. According to the principle of GBDT [21] algorithm, it can be used for feature combination and feature selection. When it is used for feature selection, the global importance of feature j is also measured by the importance of feature j in each tree. The global importance of feature j is calculated as follows:

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{J}_j^2(T_m) \tag{2}$$

where M represents the amount of decision trees.

The importance of features in a single decision tree is calculated by the following formula:

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1}\hat{i}_t^2 I(v_t = j) \tag{3}$$

where L represents the amount of leaf nodes of the decision tree; $L-1$ refers to the amount of non-leaf nodes of the decision tree; v_t is the characteristic associated with node t; $\hat{i}_t^2$ represents the reduction value of the square loss after node splitting, and I is the indicative function.

2.2. Processing Method of Unbalanced Dataset

The original dataset has a certain imbalance, which may affect the classification accuracy of the algorithm. Consequently, it is essential to balance the original dataset. In this paper, two sampling algorithms are used to increase the sampling of the minority samples, namely, synthetic minority oversampling technology (SMOTE) and adaptive synthetic (ADASYN). The SMOTE algorithm [22–26] is an improved algorithm on the basis of the random over sampling algorithm. The fundamental thought of the SMOTE algorithm is to synthesize new samples artificially by analyzing the minority samples and then adding the synthesized new samples to the dataset. The generation method of synthetic samples is as follows: (1) calculate the distance from the minority sample x to the all samples set S_{min} of the minority class, and obtain its k-nearest neighbor; (2) according to the sample's imbalance proportion to set a sampling proportion, then, several samples are randomly selected from its k-nearest neighbors for each minority sample x, assuming that the selected nearest neighbor are x_n; (3) calculate the distance from the minority sample x to each nearest neighbor x_n, denoted by $|x - x_n|$, multiply this distance by a random number between 0 and 1, and add the multiplication to sample x to produce a new sample x_{new}. This method selects a random point on the connecting line of two specific samples as the new sample. By increasing the number of minority samples, this kind of method effectively makes the minority class decision area become more common.

The calculation formula is as follows:

$$x_{new} = x + \text{rand}(0,1) * |x - x_n| \tag{4}$$

Different from the SMOTE algorithm which generates the equal number of synthetic samples for every minority class data example, the key idea of the ADASYN [26–28]

algorithm is to determine how many new samples should be resampled for every minority class sample automatically, by using density distribution as the standard. The dataset generated by the ADASYN algorithm will not only show the balanced distribution of data, but also require the classification algorithm to devote more attention to those samples that are hard to learn [27].

2.3. Classification Algorithm

This paper applies two classification models, namely, Random Forest and XGBoost. The Random Forest algorithm, introduced by Breiman, is a highly effective and most frequently used model, which can be used for classification and regression problems at the same time [29,30]. It belongs to an ensemble learning technology based on bagging. Its basic idea is to train a set of base classifiers, usually a decision tree, and then aggregate the results of the base classifiers by hard voting or weighted voting to obtain the final prediction output. Therefore, Random Forest usually performs better than a single classifier. In addition, to improve the performance of Random Forest, some strategies need to be adopted, such as the introduction of a greater randomness which can make base classifiers as independent as possible during the process of creating forests. In view of these superiorities, the Random Forest algorithm has been widely used in disease prediction and system development.

XGBoost is an optimized implementation of gradient boosting. Different from Random Forest, the base classifier of XGBoost is interrelated, and the base classifier of the latter is generated based on the former. Specifically, the latter base classifier fits the prediction residuals of the previous base classifier. Based on this integrated strategy, machine learning techniques have shown high performance in solving various disease prediction and risk stratification tasks in recent years [31–34].

3. Experiments and Results

3.1. Experimental Dataset

The Z-Alizadeh Sani dataset downloaded from UCI Machine Learning Repository consists of the medical records of 303 patients who visited Shaheed Rajaei Hospital due to chest pain. Each record contains 54 features. According to medical knowledge, each feature is the indicator of CAD diagnosis, that is, each feature is the relevant feature of CAD prediction. In the medical literature, these features can be separated into four categories, namely: demographic; symptom and examination; ECG; and laboratory and echo features. The specific information about the features of the Z-Alizadeh Sani dataset is shown in Table 1. The 303 samples of the Z-Alizadeh Sani dataset can be divided into two classes: CAD patient class and normal class. When the diameter of at least one of the three arteries is narrowed by more than or equal to 50%, the patient will be classified as CAD; otherwise, it will be considered normal [15].

3.2. Evaluation Metrics

3.2.1. Confusion Matrix

The confusion matrix is a comprehensive evaluation index system used for describing the classifier's performance. In the confusion matrix, the rows represent the real classes $y^{(i)}$ and the columns represent the predicted classes $\widehat{y}\left(x^{(i)}\right)$. In accordance with Table 2, where TP = true positive, i.e., positive instances that are actually CAD class and also correctly predicted as CAD, FP = false positive, i.e., negative instances that are actually normal class but mistakenly predicted as CAD, FN = false negative, i.e., positive instances that are actually CAD class but mistakenly predicted as normal class, TN = true negative, i.e., negative instances that are actually normal class and also correctly predicted as normal class.

Table 1. Feature name, range value and attribute type of each feature in the Z-Alizadeh Sani dataset.

Category	Feature Name	Range	Type
Demographic features	Age	30–86	continuous
	Weight	48–120	continuous
	Length	140–188	continuous
	Sex	Male, Female	categorical
	BMI	18.12–40.90	continuous
	DM	0, 1	categorical
	HTN	0, 1	categorical
	Current smoker	0, 1	categorical
	Ex-smoker	0, 1	categorical
	FH	0, 1	categorical
	Obesity (Yes (BMI > 25), else No)	Y, N	categorical
	CRF	Y, N	categorical
	CVA	Y, N	categorical
	Airway disease	Y, N	categorical
	Thyroid disease	Y, N	categorical
	CHF	Y, N	categorical
	DLP	Y, N	categorical
Symptoms and Physical examination	BP	90.0–190.0	continuous
	PR	50.0–110.0	continuous
	Edema	0, 1	categorical
	Weak peripheral pulse	Y, N	categorical
	Lung rales	Y, N	categorical
	Systolic murmur	Y, N	categorical
	Diastolic murmur	Y, N	categorical
	Typical chest pain	0, 1	categorical
	Dyspnea	Y, N	categorical
	Function class	1–4	categorical
	Atypical	Y, N	categorical
	Nonanginal	Y, N	categorical
	Exertional CP	N	categorical
	LowTH Ang	Y, N	categorical
Electrocardiography	Q Wave	0, 1	categorical
	St elevation	0, 1	categorical
	St depression	0, 1	categorical
	T inversion	0, 1	categorical
	LVH	Y, N	categorical
	Poor R progression	Y, N	categorical
	BBB	N, LBBB, RBBB	categorical
Laboratory Tests and Echocardiography	FBS	62.0–400.0	continuous
	CR	0.5–2.2	continuous
	TG	37.0–1050.0	continuous
	LDL	18.0–232.0	continuous
	HDL	15.9–111.0	continuous
	BUN	6.0–52.0	continuous
	ESR	1–90	continuous
	HB	8.9–17.6	continuous
	K	3.0–6.6	continuous
	Na	128.0–156.0	continuous
	WBC	3700–18,000	continuous
	Lymph	7.0–60.0	continuous
	Neut	32.0–89.0	continuous
	PLT	25.0–742.0	continuous
	EF-TTE	15.0–60.0	continuous
	Region RWMA	0, 1, 2, 3, 4	categorical
	VHD	Mild, N, moderate, severe	categorical
Cath_label	Cath		

Table 2. Confusion matrix.

	$\widehat{y}\left(x^{(i)}\right) = 0$	$\widehat{y}\left(x^{(i)}\right) = 1$
$\widehat{y}(i) = 0$	TN	FP
$\widehat{y}(i) = 1$	FN	TP

3.2.2. Classification Metrics

The original dataset has a certain imbalance. Only using accuracy cannot measure the models' performance effectively. Thus, besides the accuracy, we also calculated other model evaluation metrics such as recall, specificity, precision, F_1 score and AUC, to evaluate the performance of the classifier models.

1. Accuracy

Accuracy measures all the samples that are predicted correctly, including positive samples and negative samples.

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

Accuracy is an evaluation index that is regularly used and easily understood. The influence of positive and negative samples on accuracy is the same. However, in the medical domain, doctors and patients actually pay more attention to the positive samples, namely the CAD samples. At this time, the costs brought on by the missed diagnosis of positive samples and misdiagnosis of negative samples are different. In these circumstances, only using accuracy to assess the performance of a classifier is insufficient.

2. Precision

Precision is used to measure the proportion of true positive samples in instances that are predicted to be positive.

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

From the perspective of a positive sample, precision tends to measure how likely it is that an instance predicted to be a positive sample is indeed a true positive sample.

3. Recall

Recall represents the proportion of samples that are correctly predicted in all positive samples. Recall is an important evaluation index that measures the classifier's ability to recognize positive samples.

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

4. F_1 score

Precision and recall often restrict each other. Therefore, the F_1 score is introduced, that is the weighted harmonic average of recall and precision. A higher F_1 score indicates that the test method is more effective.

$$\text{F}_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

5. Specificity

Specificity represents the ratio of samples that are correctly predicted in all negative samples.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

6. AUC

The receiver operating characteristics (ROC) curve chart with true positive rate (TPR) as the y-axis and false positive rate (FPR) as the x-axis is the relationship diagram between true positive rate and false positive rate, which actually reflects the relationship between specificity and recall. The receiver operating characteristics (ROC) curve contributes significantly to visually display the power of the classifier model's classification capability. The ROC curve of the classification model for CAD prediction is closer to the upper left corner, indicating that the prediction performance of the classifier for CAD is stronger. The area under the ROC curve, namely, area under curve (AUC), is usually used to quantitatively measure the ROC curve. In other words, the closer the AUC is to 1, the more accurate the classifier is in the prediction of CAD, that is, the prediction performance of the classifier is higher.

3.3. Experimental Results

In this part, the experimental results of the XGBoost algorithm and Random Forest algorithm combining four feature processing technologies and two dataset balancing methods is reported. Firstly, four feature processing technologies, namely feature smoothing, feature encoding, feature construction and feature selection, are applied to the original dataset. A total of five sets of data are obtained adding the original dataset after the above handle. Meanwhile, two dataset balancing methods, synthetic minority oversampling technique (SMOTE) and adaptive synthetic (ADASYN), are applied to balance the classes in the datasets, respectively. In total, we have 15 sets of data. Furthermore, the performances of the XGBoost algorithm and Random Forest algorithm are evaluated on these datasets separately. A 10-fold cross-validation technology is also used for model development.

3.3.1. Results Obtained on Original Dataset and Two Balanced Datasets

The classification performance results of the XGBoost algorithm and Random Forest algorithm for CAD prediction in terms of accuracy, recall, precision, F_1 score, specificity and AUC on three datasets which include the original dataset and two datasets balanced by SMOTE and ADASYN, respectively, are reported in this section. The original dataset has a certain imbalance, which shows that among 303 samples, 216 samples are CAD (accounting for 71.29%) and 87 samples are normal (accounting for 28.71%). The dataset balanced by the SMOTE method contains 432 samples. At this point, the number of samples of CAD class and normal class are equal, that is, each class consists of 216 samples. However, the dataset balanced using ADASYN includes 426 samples, among them, 216 samples are CAD, and 210 samples are normal. The average testing result in terms of accuracy, precision, recall, F_1 score, specificity and AUC obtained by the XGBoost algorithm and Random Forest algorithm for the three datasets are reported in Table 3.

Table 3. The average testing results obtained on the original dataset and two datasets balanced by SMOTE and ADASYN respectively.

Datasets	Algorithms	Recall	F_1	Accuracy	Precision	Specificity	AUC
Original data	Random Forest	0.909 ± 0.051	0.923 ± 0.038	0.887 ± 0.056	0.940 ± 0.051	0.834 ± 0.117	0.92 ± 0.05
	XGBoost	0.909 ± 0.072	0.929 ± 0.044	0.894 ± 0.068	0.954 ± 0.046	0.856 ± 0.106	0.93 ± 0.05
Balanced data with SMOTE	Random Forest	0.939 ± 0.086	0.941 ± 0.045	0.938 ± 0.052	0.949 ± 0.044	0.936 ± 0.041	0.98 ± 0.03
	XGBoost	0.940 ± 0.087	0.943 ± 0.044	0.940 ± 0.052	0.953 ± 0.037	0.940 ± 0.035	0.97 ± 0.03
Balanced data with ADASYN	Random Forest	0.937 ± 0.081	0.930 ± 0.040	0.928 ±0.045	0.930 ± 0.048	0.918 ± 0.041	0.96 ± 0.04
	XGBoost	0.939 ± 0.082	0.942 ± 0.042	0.939 ± 0.048	0.953 ± 0.048	0.941 ± 0.096	0.97 ± 0.03

The experimental results show that the XGBoost model on the dataset balanced by the SMOTE method achieves the best performance with a classification accuracy of 94.0%,

F_1 score of 94.3%, recall of 94.0%, precision of 95.3%, specificity of 94.0% and AUC of 0.97. From the result, it can be inferred that two dataset balancing methods can enhance the capability of the XGBoost and Random Forest model for predicting CAD. The best results appear in the combination of the XGBoost classification model with the SMOTE method. In addition, it can be found that the XGBoost algorithm performs better than the Random Forest algorithm on the datasets used in this section.

3.3.2. Results Obtained on Datasets Processed by Feature Smoothing and Two Dataset Balancing Methods

The performance results of the XGBoost algorithm and Random Forest algorithm for CAD prediction in respect of classification accuracy, recall, precision, F_1 score, specificity and AUC for the datasets processed by feature smoothing technology and two dataset balancing methods are discussed in this section. Feature smoothing technology only processes the outliers in the dataset and does not change the size of the dataset. Therefore, the dataset processed by feature smoothing technology still has a certain imbalance, in that among 303 samples, 216 samples are CAD and 87 samples are normal. The dataset balanced by the SMOTE method contains 432 samples, of which the numbers of samples classified as CAD and normal are both 216. The dataset balanced using the ADASYN method consists of 427 samples, among them, the numbers of samples classified as CAD and normal are 216 and 211 respectively. The average testing result in terms of accuracy, precision, recall, F_1 score, specificity and AUC obtained by the XGBoost algorithm and Random Forest algorithm for the three datasets are reported in Table 4.

Table 4. The average testing result obtained on datasets processed by feature smoothing and two dataset balancing methods.

Datasets	Algorithms	Recall	F_1	Accuracy	Precision	Specificity	AUC
Data processed by feature smoothing	Random Forest	0.892 ± 0.051	0.922 ± 0.033	0.884 ± 0.053	0.958 ± 0.039	0.866 ± 0.094	0.91 ± 0.06
	XGBoost	0.914 ± 0.070	0.931 ± 0.037	0.898 ± 0.059	0.954 ± 0.046	0.857 ± 0.100	0.93 ± 0.06
Data processed by feature smoothing and SMOTE	Random Forest	0.939 ± 0.081	0.933 ± 0.046	0.931 ± 0.051	0.931 ± 0.031	0.923 ± 0.033	0.98 ± 0.02
	XGBoost	0.945 ± 0.083	0.935 ± 0.046	0.933 ± 0.053	0.930 ± 0.038	0.921 ± 0.040	0.98 ± 0.02
Data processed by feature smoothing and ADASYN	Random Forest	0.936 ± 0.087	0.936 ± 0.052	0.932 ± 0.059	0.940 ± 0.036	0.928 ± 0.041	0.98 ± 0.02
	XGBoost	0.941 ± 0.078	0.942 ± 0.041	0.939 ± 0.045	0.948 ± 0.045	0.937 ± 0.040	0.97 ± 0.03

All the experimental results show that, compared with Table 3, the performance results of the XGBoost model and the Random Forest model for the datasets processed by feature smoothing are generally reduced. In detail, the feature smoothing technology only improves the performance results of the XGBoost algorithm on the original dataset and the performance results of the Random Forest algorithm on the dataset balanced by ADASYN. In addition, the classification performance of the two models on other datasets is all degraded.

Moreover, it can be found from the experimental results that the performance of the XGBoost algorithm is still better than the Random Forest algorithm on the datasets in this section. The two dataset balancing methods still have the ability to improve the prediction performance of the model. Different to the original dataset, the best result on the dataset processed by feature smoothing technology comes from the combination of the ADASYN method and the XGBoost algorithm. The best recall, F_1 score, accuracy, precision, specificity and AUC are 94.1%, 94.2%, 93.9%, 94.8%, 93.7% and 0.97 respectively.

3.3.3. Results Obtained on Datasets Processed by Feature Encoding and Two Dataset Balancing Methods

The performance results of the XGBoost algorithm and Random Forest algorithm for CAD prediction in respect of classification accuracy, recall, precision, F_1 score, specificity and AUC for the datasets processed by feature encoding technology and two dataset balancing methods are discussed in this section. The dataset processed by feature encoding

technology contains 216 CAD samples and 87 normal samples. The dataset balanced by the SMOTE method consists of 432 samples, of which 216 samples are CAD and 216 samples are normal. The dataset balanced using the ADASYN method includes 421 samples, among them, 216 samples are CAD and 205 samples are normal. The average testing result in terms of accuracy, precision, recall, F_1 score, specificity and AUC obtained by the XGBoost algorithm and Random Forest algorithm for the three datasets are reported in Table 5.

Table 5. The average testing result obtained on datasets processed by feature encoding and two dataset balancing methods.

Datasets	Algorithms	Recall	F_1	Accuracy	Precision	Specificity	AUC
Data processed by feature encoding	Random Forest	0.900 ± 0.058	0.920 ± 0.027	0.881 ± 0.046	0.944 ± 0.041	0.833 ± 0.070	0.91 ± 0.06
	XGBoost	0.918 ± 0.053	0.925 ± 0.032	0.891 ± 0.047	0.935 ± 0.047	0.824 ± 0.101	0.93 ± 0.06
Data processed by feature encoding and SMOTE	Random Forest	0.949 ± 0.071	0.933 ± 0.035	0.933 ± 0.037	0.925 ± 0.065	0.917 ± 0.056	0.98 ± 0.01
	XGBoost	0.943 ± 0.075	0.932 ± 0.037	0.931 ± 0.041	0.926 ± 0.037	0.918 ± 0.030	0.97 ± 0.02
Data processed by feature encoding and ADASYN	Random Forest	0.924 ± 0.081	0.926 ± 0.040	0.922 ± 0.047	0.935 ± 0.043	0.919 ± 0.038	0.96 ± 0.03
	XGBoost	0.933 ± 0.084	0.928 ± 0.042	0.924 ± 0.048	0.930 ± 0.043	0.915 ± 0.039	0.97 ± 0.03

It can be seen from the experimental results that, compared with Tables 3 and 4, the performance results of the XGBoost and the Random Forest model in terms of recall for the dataset processed by feature encoding and the dataset processed by feature encoding and SMOTE are slightly improved. However, all performance results of the two models on other datasets are degraded.

It should be noted that in this section that the performance results of the Random Forest model on the dataset processed by feature encoding and SMOTE is better than the XGBoost model. Additionally, different to the above two parts, on the dataset processed by feature encoding technology, the best result comes from the combination of the SMOTE method and Random Forest model. The best recall, F_1 score, accuracy, precision, specificity and AUC are 94.9%, 93.3%, 93.3%, 92.5%, 91.7% and 0.98 respectively. In the same way, it can be found from the experimental results that the two dataset balancing methods still have the ability to improve the prediction performance of the model.

3.3.4. Results Obtained on Datasets Processed by Feature Selection and Two Dataset Balancing Methods

The performance results of the XGBoost algorithm and Random Forest algorithm for CAD prediction in respect of classification accuracy, recall, precision, F_1 score, specificity and AUC for the datasets processed by feature selection technology and two dataset balancing methods are discussed in this section. Twelve features that are considered to be the most relevant features of CAD prediction are selected by feature selection technology based on the GBDT algorithm. The details of the 12 selected features are shown in Table 6 and Figure 4. In Table 6 the names of the 12 important features and the corresponding feature importance are shown. Figure 4 expresses the information in Table 6 visually. As can be seen from Table 6 and Figure 4, the 12 selected features are typical chest pain, age, TG, region RWMA, BMI, EF-TTE, ESR, weight, HTN, LDL, nonanginal and T inversion. The significant correlation and indicative relationship between these 12 medical features and CAD diagnosis have also been confirmed by medical experts and medical literature [4]. This shows the effectiveness of our feature selection method.

Table 6. 12 Selected features.

Feature_Name	Importance [1]
Typical chest pain	0.329970
Age	0.170756
TG	0.077998
Region RWMA	0.076567
BMI	0.068513
EF-TTE	0.066505
ESR	0.044726
Weight	0.040609
HTN	0.039968
LDL	0.037006
Nonanginal	0.026549
T inversion	0.020833

[1] The global importance of feature *j* calculated by GBDT. TG, triglyceride. RWMA, regional wall motion abnormalities. EF–TTE, ejection fraction–transthoracic echocardiography. ESR, erythrocyte sedimentation rate. HTN, hypertension. LDL, low density lipoprotein.

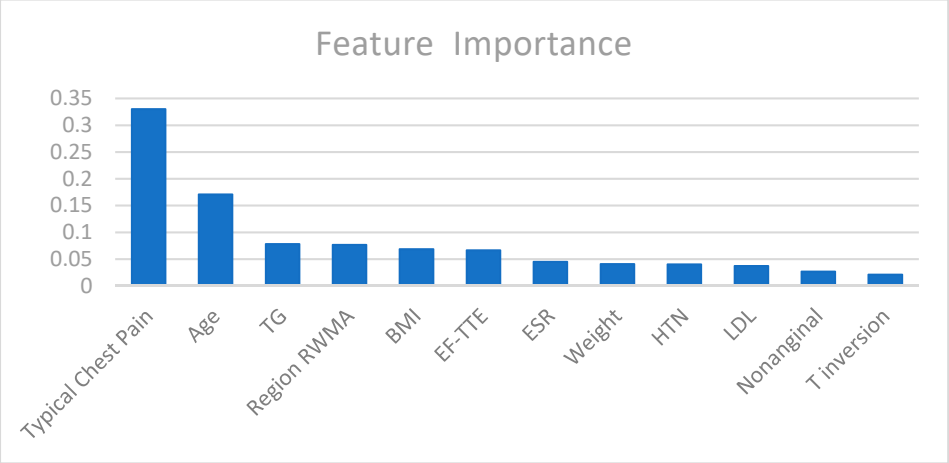

Figure 4. Feature importance of 12 selected features.

The dataset processed by feature selection technology contains 216 CAD samples and 87 normal samples. The dataset balanced by SMOTE method consists of 432 samples, of which 216 samples are CAD and 216 samples are normal. The dataset balanced using the ADASYN method includes 428 samples, and among them, CAD samples and normal samples are 216 and 212, respectively. The average testing result in terms of accuracy, precision, recall, F_1 score, specificity and AUC obtained by the XGBoost algorithm and Random Forest algorithm for the three datasets are reported in Table 7.

Table 7. The average testing result obtained on datasets processed by feature selection and two dataset balancing methods.

Datasets	Algorithms	Recall	F_1	Accuracy	Precision	Specificity	AUC
Data processed by feature selection	Random Forest	0.899 ± 0.067	0.930 ± 0.039	0.894 ± 0.062	0.968 ± 0.036	0.884 ± 0.091	0.95 ± 0.04
	XGBoost	0.927 ± 0.058	0.939 ± 0.039	0.911 ± 0.057	0.954 ± 0.054	0.870 ± 0.127	0.94 ± 0.05
Data processed by feature selection and SMOTE	Random Forest	0.944 ± 0.075	0.937 ± 0.037	0.935 ± 0.042	0.935 ± 0.043	0.927 ± 0.040	0.97 ± 0.03
	XGBoost	0.943 ± 0.079	0.942 ± 0.052	0.940 ± 0.056	0.944 ± 0.046	0.937 ± 0.050	0.96 ± 0.04
Data processed by feature selection and ADASYN	Random Forest	0.944 ± 0.073	0.941 ± 0.038	0.939 ± 0.042	0.944 ± 0.051	0.935 ± 0.047	0.97 ± 0.02
	XGBoost	0.946 ± 0.067	0.943 ± 0.040	0.942 ± 0.043	0.944 ± 0.041	0.938 ± 0.042	0.97 ± 0.04

From the result of Table 7, it can be inferred that the feature selection technology based on the GBDT algorithm can improve the performance of the XGBoost model and Random

Forest model for predicting CAD. The performance results of the XGBoost algorithm and Random Forest algorithm on the datasets processed by feature selection technology based on the GBDT algorithm are significantly better than the performance results of two models on the datasets processed by feature smoothing technology and feature encoding technology. At the same time, when compared with Table 3, it can be seen that the feature selection technology based on the GBDT algorithm can improve the performance results of the XGBoost model and Random Forest model on the original dataset and the balanced dataset by the ADASYN method. However, on the balanced dataset by the SMOTE method, the feature selection technology based on the GBDT algorithm only promotes the performance of the XGBoost model and Random Forest model in terms of recall. Therefore, it can be concluded that the feature selection technology based on the GBDT algorithm can improve the performance of the models for CAD prediction by identifying the most relevant, important and less redundant features. In this section, the best results appear in the combination of the XGBoost classification model with the ADASYN method with a classification accuracy of 94.2%, F_1 score of 94.3%, recall of 94.6%, precision of 94.4%, specificity of 93.8% and AUC of 0.97. In addition, it also can be found that the XGBoost algorithm performs better than the Random Forest algorithm on the datasets used in this section.

3.3.5. Results Obtained on Datasets Processed by Feature Construction and Two Dataset Balancing Methods

The performance results of the XGBoost algorithm and Random Forest algorithm for prediction of CAD in terms of classification accuracy, precision, recall, F_1 score, specificity and AUC for the datasets processed by feature construction technology and two dataset balancing methods are discussed in this section. Feature construction technology increases the feature dimension of the samples to 120 dimensions without changing the size of the samples. Therefore, the dataset processed by feature construction technology still contains 303 samples, of which 216 samples are CAD and 87 samples are normal. The dataset balanced by the SMOTE method consists of 432 samples, out of which the CAD and normal samples are both 216. The dataset balanced using ADASYN method includes 420 samples, among them, 216 samples are CAD and 204 samples are normal. The average testing result in terms of accuracy, precision, recall, F_1 score, specificity and AUC obtained by the XGBoost algorithm and Random Forest algorithm for the three datasets are reported in Table 8.

Table 8. The average testing result obtained on datasets processed by feature construction and two dataset balancing methods.

Datasets	Algorithms	Recall	F_1	Accuracy	Precision	Specificity	AUC
Data processed by feature construction	Random Forest	0.897 ± 0.072	0.918 ± 0.040	0.878 ± 0.064	0.944 ± 0.035	0.829 ± 0.089	0.90 ± 0.06
	XGBoost	0.911 ± 0.073	0.929 ± 0.036	0.895 ± 0.059	0.954 ± 0.041	0.854 ± 0.090	0.93 ± 0.05
Data processed by feature construction and SMOTE	Random Forest	0.933 ± 0.085	0.934 ± 0.046	0.931 ± 0.051	0.940 ± 0.037	0.929 ± 0.035	0.97 ± 0.03
	XGBoost	0.961 ± 0.048	0.946 ± 0.030	0.947 ± 0.029	0.934 ± 0.053	0.932 ± 0.049	0.98 ± 0.02
Data processed by feature construction and ADASYN	Random Forest	0.933 ± 0.088	0.921 ± 0.048	0.917 ± 0.057	0.916 ± 0.051	0.899 ± 0.052	0.96 ± 0.04
	XGBoost	0.946 ± 0.065	0.938 ± 0.037	0.936 ± 0.041	0.935 ± 0.043	0.925 ± 0.044	0.98 ± 0.02

It can be seen from the experimental results that the best performance results of the classification models for prediction of CAD in terms of classification accuracy, recall and F_1 score are obtained by the XGBoost model which is trained on the dataset processed by feature construction and SMOTE method. The best recall, F_1 score, accuracy, precision, specificity and AUC are 96.1%, 94.6%, 94.7%, 93.4%, 93.2% and 0.98 respectively. From the result, it can be inferred that the XGBoost model combining feature construction technology and SMOTE method has significant ability to identify CAD patients. The higher F_1 score and AUC show that the performance of the model is stable and effective.

4. Discussion

Feature engineering means a range of technologies that use a sequence of engineering methods to filter out more relevant features from the raw data to promote the training result of the classifier. The function of feature engineering is to remove redundant features and noises existing in the raw data, and select or construct features that can more effectively describe the relationship between the problem to be solved, and the prediction model. In our work, to explore the method of detecting CAD quickly and accurately, the performance of the XGBoost algorithm and Random Forest algorithm is evaluated on datasets processed by four feature processing technologies and two dataset balancing methods. The four feature processing techniques are feature smoothing, feature encoding, feature selection and feature construction. The two dataset balancing methods are based on the SMOTE algorithm and ADASYN algorithm, respectively. Moreover, 10-fold cross-validation technology is used to test the stability and accuracy of the model. Experimental results demonstrate that the four feature processing technologies have different effects on the performance of the classification model on different datasets. Among them, the impact of feature construction technology on the performance of the classification model is prominent. Figure 5a–d shows the effects of four feature processing technologies and two datasets balancing methods on the performance of XGBoost algorithm in terms of accuracy, recall, F_1 score and AUC, respectively. It can be found from Figure 5 that on the dataset processed by feature construction technology and the SMOTE algorithm, the XGBoost classification model produces the best performance results in terms of classification accuracy, recall, F_1 score and AUC. At this time, the XGBoost classification model has the strongest recognition ability for positive samples. Secondly, the feature selection technology based on the GBDT algorithm also significantly improves the classification model. On the dataset processed by feature selection technology and the ADASYN algorithm, the XGBoost classification model also achieves better classification performance. It is worth noting that the better performance results are achieved based on 12 features selected by the GBDT algorithm. However, it is regrettable that feature smoothing technology and feature encoding technology have a poor effect on enhancing the capability of the classification model. Furthermore, it can also be seen from Figure 5 that the two dataset balancing methods can significantly improve the performance of the classification model.

In addition, the comparison of the performance results of our proposed method with the performance results of previous studies on the Z-Alizadeh Sani dataset reported in the literature is shown in Table 9. It can be seen from Table 9 that the proposed method has achieved better performance than existing research. It should be noted that there are many values marked as NR, which represents that these classification metrics have not been reported in the literature, in Table 9. However, these metrics are vital to evaluate the performance of medical models, especially to evaluate the performance of classification models that are trained on imbalanced datasets. Additionally, in Table 9, the column name "FeatureNums" refers to the number of features used for model training and testing. According to the column "FeatureNums", feature selection technology has been applied to almost all studies reported in the literature. In our study, 12 features that are considered to be most relevant to CAD prediction have been selected by the GBDT algorithm, and the performance results of the classification model trained on these 12 features, which is the least number of features in the comparison literature, are very promising. However, although some performance results reported in the literature [16,17] are better than our study, the following points need to be observed: (1) the accuracy and recall in [16] were obtained based on 500 samples; (2) the number of features used in [16,17] were more than used in our study; (3) our proposed method also achieves very competitive results in terms of specificity, precision, F_1 score and AUC, especially the best AUC. Overall, the experimental results clearly demonstrate the robustness and stability of our proposed method in CAD diagnosis and prediction, and it can be seen from Table 9 that our proposed method provides better results when compared with other studies that already exist in the literature.

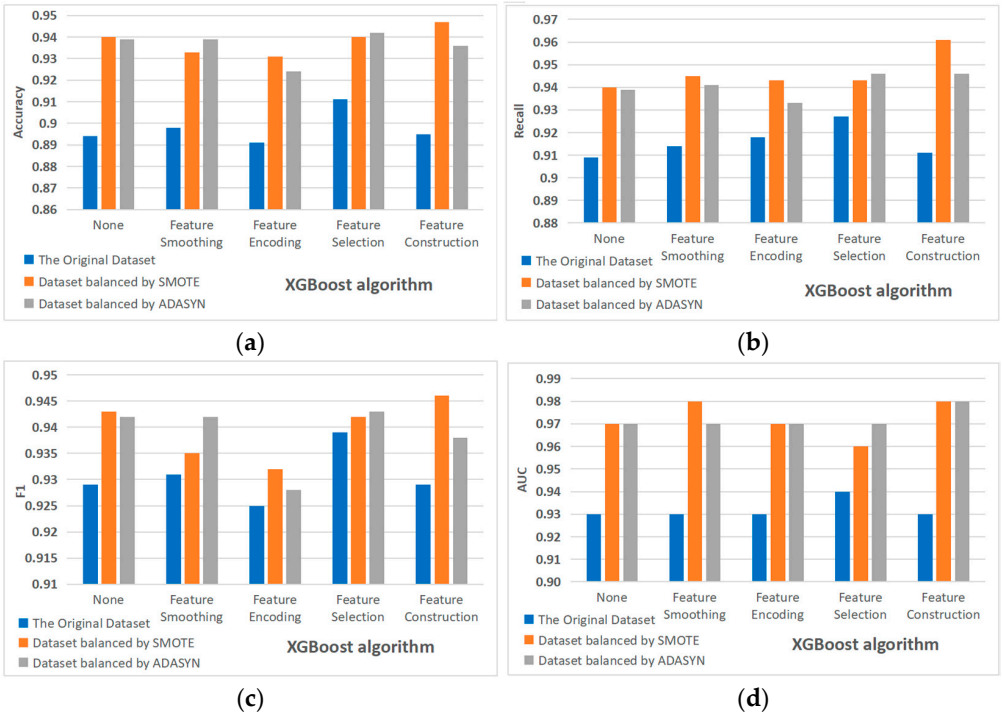

Figure 5. Comparison of the effects of four feature processing technologies and two dataset balancing methods on the performance of XGBoost algorithm. (**a**) In terms of accuracy, (**b**) In terms of recall, (**c**) In terms of F_1, (**d**) In terms of AUC.

Table 9. Comparison of the performance results of our proposed method with the performance results of previous studies on the Z-Alizadeh Sani dataset.

Method	Feature Nums	Accuracy %	Recall %	Specificity %	Precision %	F_1%	AUC
SMO [35]	34	92.09	97.22	79.31	NR	NR	NR
SMO + information gain [15]	33	94.08	96.30	88.51	NR	NR	NR
KNN (K1 KNN) [36]	NR	90.91	93.33	85.71	93.33	93.33	NR
NN + genetic [13]	22	93.85	97	92	NR	NR	NR
NB + genetic algorithm [37]	32	88.16	88.00	87.78	NR	NR	NR
Ensemble [38]	25	86.49	73.61	91.67	NR	0.75	0.83
SVM + feature engineering [16]	28	96.4	100	88.1	NR	NR	0.92
NE-nu-SVC [17]	16	94.66	94.70	NR	94.70	94.70	0.966
N2GC-nuSVM [10]	29	93.08	NR	NR	NR	91.51	NR
XGBoost + hybrid FSA + FA + ETCA + SMOTE [9]	27	92.58	92.99	NR	92.59	90.62	NR
Hybrid PSO-EmNN coupled with feature selection [39]	22	88.34	91.85	78.98	92.37	92.12	NR
XGBoost + GDBT + ADASYN *	12	94.2	94.6	93.8	94.4	94.3	0.97
XGBoost + feature construction + SMOTE *	120	94.7	96.1	93.2	93.4	94.6	0.98

* Our proposed methods.

In consideration of the above, the major advantages of our proposed method are as follows: (1) application of the XGBoost algorithm to the Z-Alizadeh Sani dataset for earlier and effective diagnosis of CAD; (2) a series of feature processing techniques, such as feature smoothing, feature frequency encoding, feature construction and feature selection technology, were applied to the Z-Alizadeh Sani dataset to reduce feature redundancy and improve the accuracy of classification models for CAD prediction; (3) application of the feature selection method based on GBDT algorithm on the Z-Alizadeh Sani dataset; (4) two

classical datasets balancing methods were applied to the Z-Alizadeh Sani dataset to solve the problem of dataset imbalance; and (5) classification metrics such as accuracy, recall, specificity, precision, F_1 score and AUC were used to validate the model performance. Of course, our research also has some shortcomings, such as: (1) the dataset used was small; (2) more ensemble learning techniques were not tried in this research. This will be the direction of future work.

5. Conclusions

CAD is one of the diseases with the highest morbidity and mortality in the world. The goal of how to achieve rapid and accurate CAD detection is being pursued by many researchers, scholars and doctors around the world. In this study, four different feature processing techniques, including feature smoothing, feature encoding, feature construction and feature selection, were applied to the Z-Alizadeh Sani dataset to explore methods that can improve the performance of classification models for CAD detection. We used the XGBoost algorithm and Random Forest algorithm as classifiers and applied a 10-fold cross-validation technique to test the stability of the model. SMOTE algorithm and ADASYN algorithm were used to balance the imbalanced dataset. Model evaluation measurements such as accuracy, recall, specificity, precision, F_1 score and AUC were used to evaluate the performance of the classification model. Experimental results show that, compared with the most advanced algorithms in the literature, our method is very competitive and can be used by medical staff for clinical auxiliary diagnosis.

Author Contributions: Conceptualization, Y.Y., Z.L. and X.W.; methodology, S.Z., Y.Y. and Z.Y.; software, S.Z.; validation, S.Z. and Z.Y.; formal analysis, S.Z., Z.L. and X.W.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z. and Y.Y.; supervision, Z.L. and X.W.; funding acquisition, Z.L. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by two funds: Basic Research of the Ministry of Science and Technology, China (2013FY114000), and National Key R&D Program of China (2016YFF0201003).

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

References

1. Mensah, G.A.; Roth, G.A.; Fuster, V. The Global Burden of Cardiovascular Diseases and Risk Factors 2020 and Beyond. *JACC* **2019**, *74*, 2529–2532. [CrossRef] [PubMed]
2. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1223–1249. [CrossRef]
3. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1204–1222. [CrossRef]
4. Zipes, D.P.; Libby, P.; Bonow, R.O. *Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2018.
5. Jayaraman, V.; Sultana, H.P. Artifificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. *J. Ambient Intell. Humaniz. Comput.* **2019**, 1–10. [CrossRef]
6. Liu, M.; Kim, Y. Classification of Heart Diseases Based on ECG Signals Using Long Short-Term Memory. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 2707–2710.
7. Vijayashree, J.; Sultana, H.P. Heart disease classification using hybridized ruzzo-tompa memetic based deep trained neocognitron neural network. *Health Technol.* **2019**, *10*, 207–216. [CrossRef]
8. Alizadehsani, R.; Abdar, M.; Roshanzamir, M.; Khosravi, A.; Kebria, P.M.; Khozeimeh, F.; Nahavandi, S.; Sarrafzadegan, N.; Acharya, U.R. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Comput. Biol. Med.* **2019**, *111*, 103346. [CrossRef]
9. Nasarian, E.; Abdar, M.; Fahami, M.A.; Alizadehsani, R.; Hussaind, S.; Basiri, M.E.; Zomorodi-Moghadam, M.; Zhou, X.J.; Pławiak, P.; Acharya, U.R.; et al. Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognit. Lett.* **2020**, *133*, 33–40. [CrossRef]
10. 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]

11. Kolukisa, B.; Hacilar, H.; Goy, G.; Kus, M.; Bakir-Gungor, B.; Aral, A.; Gungor, V.C. Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2232–2238.
12. Zomorodi-moghadam, M.; Abdar, M.; Davarzani, Z.; Zhou, X.J.; Pławiak, P.; Acharya, U.R. Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease. *Expert Syst.* **2019**, *38*, e12485. [CrossRef]
13. Arabasadi, Z.; Alizadehsani, R.; Roshanzamir, M.; Moosaei, H.; Yarifard, A.A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Comput. Methods Programs Biomed.* **2017**, *141*, 19–26. [CrossRef]
14. Joloudari, J.H.; Joloudari, E.H.; 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]
15. Alizadehsani, R.; Habibi, J.; Hosseini, M.J.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B.; Sani, Z.A. A data mining approach for diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* **2013**, *111*, 52–61. [CrossRef]
16. Alizadehsani, R.; Hosseini, M.J.; Khosravi, A.; Khozeimeh, F.; Roshanzamir, M.; Sarrafzadegan, N.; Nahavandi, S. Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. *Comput. Methods Programs Biomed.* **2018**, *162*, 119–127. [CrossRef]
17. Abdar, M.; Acharya, U.R.; Sarrafzadegan, N.; Makarenkov, V. Ne-nu-svc: A new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease. *IEEE Access* **2019**, *7*, 167605–167620. [CrossRef]
18. Ashish, L.; Kumar, S.; Yeligeti, S. Ischemic heart disease detection using support vector machine and extreme gradient boosting method. *Mater. Today Proc.* **2021**. [CrossRef]
19. Tian, Z.; Chen, C.Y.; Fan, Y.M.; Ou, X.J.; Wang, J.; Ma, X.L.; Xu, J.G. Glioblastoma and Anaplastic Astrocytoma: Differentiation Using MRI Texture Analysis. *Front. Oncol.* **2019**, *9*, 876. [CrossRef]
20. Qing, Y.; Zhi, J.C.; Ying, L.T. Prediction of aptamer–protein interacting pairs based on sparse autoencoder feature extraction and an ensemble classifier. *Math. Biosci.* **2019**, *311*, 103–108. [CrossRef]
21. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
22. Liu, T.; Moore, A.W.; Gray, A.; Yang, K. An investigation of practical approximate nearest neighbor algorithms. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2004; pp. 825–832.
23. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
24. Xu, X.L.; Chen, W.; Sun, Y.F. Over-sampling algorithm for imbalanced data classification. *J. Syst. Eng. Electron.* **2019**, *30*, 1182–1191. [CrossRef]
25. Lee, H.S.; Jung, S.; Kim, M.; Kim, S. Synthetic Minority Over-Sampling Technique based on Fuzzy C-means Clustering for Imbalanced Data. In Proceedings of the 2017 International Conference on Fuzzy Theory and Its Applications (iFUZZY), Taiwan, China, 12–15 November 2017. [CrossRef]
26. Gosain, A.; Sardana, S. Handling Class Imbalance Problem using Oversampling Techniques: A Review. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 79–85. [CrossRef]
27. He, H.; Bai, Y.; Garcia, E.A.; Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1322–1328.
28. Satapathy, S.K.; Mishra, S.; Mallick, P.K.; Chae, G.S. ADASYN and ABC-optimized RBF convergence network for classification of electroencephalograph signal. *Pers. Ubiquitous Comput.* **2021**, 1–17. [CrossRef]
29. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
30. Akyol, K.; Çalik, E.; Bayir, Ş.; Şen, B.; Çavuşoğlu, A. Analysis of demographic characteristics creating coronary artery disease susceptibility using random forests classifier. *Procedia Comput. Sci.* **2015**, *62*, 39–46. [CrossRef]
31. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM (Association for Computing Machinery) Digital Library: New York, NY, USA, 2016; pp. 785–794.
32. Chen, Y.; Wang, X.; Jung, Y.; Abedi, V.; Zand, R.; Bikak, M.; Adibuzzaman, M. Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost. *Physiol. Meas.* **2018**, *39*, 104006. [CrossRef]
33. Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inf.* **2017**, *4*, 159. [CrossRef]
34. Van Rosendael, A.R.; Maliakal, G.; Kolli, K.K.; Beecy, A.; Al'Aref, S.J.; Dwivedi, A.; Singh, G.; Panday, M.; Kumar, A.; Ma, X.Y.; et al. Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry. *J. Cardiovasc. Comput. Tomogr.* **2018**, *12*, 204–209. [CrossRef]
35. Alizadehsani, R.; Hosseini, M.J.; Sani, Z.A.; Ghandeharioun, A.; Boghrati, R. Diagnosis of coronary artery disease using cost-sensitive algorithms. In Proceedings of the IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, 10 December 2012; pp. 9–16.

36. Dekamin, A.; Sheibatolhamdi, A. A data mining approach for coronary artery disease prediction in Iran. *J. Adv. Med. Sci. Appl. Technol.* **2017**, *3*, 29–38. [CrossRef]
37. Li, H.; Wang, X.P.; Li, Y.; Qin, C.J.; Liu, C.C. Comparison between medical knowledge based and computer automated feature selection for detection of coronary artery disease using imbalanced data. In Proceedings of the BIBE 2018, International Conference on Biological Information and Biomedical Engineering, Shanghai, China, 6–8 June 2018; pp. 1–4.
38. Cüvitoğlu, A.; Işik, Z. Classification of cad dataset by using principal component analysis and machine learning approaches. In Proceedings of the 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE), Istanbul, Turkey, 3–5 May 2018; pp. 340–343.
39. Shahid, A.H.; Singh, M.P. A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network. *Biocybern. Biomed. Eng.* **2020**, *40*, 1568–1585. [CrossRef]