

## Article

# Coronary Artery Disease Detection Model Based on Class Balancing Methods and LightGBM Algorithm

Shasha Zhang <sup>1</sup>, Yuyu Yuan <sup>1,2</sup>, Zhonghua Yao <sup>3</sup>, Jincui Yang <sup>1,2</sup>, Xinyan Wang <sup>4</sup> and Jianwei Tian <sup>4,\*</sup>

<sup>1</sup> 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.); jincuiyang@bupt.edu.cn (J.Y.)

<sup>2</sup> Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing 100876, China

<sup>3</sup> School of Computer and Engineering, Fuyang Normal University, Fuyang 236000, China; yaozh@fynu.edu.cn

<sup>4</sup> Air Force Medical Center, Chinese People's Liberation Army, Beijing 100142, China; wxy61092@163.com

\* Correspondence: tianjianwei029@aliyun.com

**Abstract:** Coronary artery disease (CAD) is a disease with high mortality and disability. By 2019, there were 197 million CAD patients in the world. Additionally, the number of disability-adjusted life years (DALYs) owing to CAD reached 182 million. It is widely known that the early and accurate diagnosis of CAD is the most efficient method to reduce the damage of CAD. In medical practice, coronary angiography is considered to be the most reliable basis for CAD diagnosis. However, unfortunately, due to the limitation of inspection equipment and expert resources, many low- and middle-income countries do not have the ability to perform coronary angiography. This has led to a large loss of life and medical burden. Therefore, many researchers expect to realize the accurate diagnosis of CAD based on conventional medical examination data with the help of machine learning and data mining technology. The goal of this study is to propose a model for early, accurate and rapid detection of CAD based on common medical test data. This model took the classical logistic regression algorithm, which is the most commonly used in medical model research as the classifier. The advantages of feature selection and feature combination of tree models were used to solve the problem of manual feature engineering in logical regression. At the same time, in order to solve the class imbalance problem in Z-Alizadeh Sani dataset, five different class balancing methods were applied to balance the dataset. In addition, according to the characteristics of the dataset, we also adopted appropriate preprocessing methods. These methods significantly improved the classification performance of logistic regression classifier in terms of accuracy, recall, precision, F1 score, specificity and AUC when used for CAD detection. The best accuracy, recall, F1 score, precision, specificity and AUC were 94.7%, 94.8%, 94.8%, 95.3%, 94.5% and 0.98, respectively. Experiments and results have confirmed that, according to common medical examination data, our proposed model can accurately identify CAD patients in the early stage of CAD. Our proposed model can be used to help clinicians make diagnostic decisions in clinical practice.

**Keywords:** cardiovascular disease; coronary artery disease; lightGBM; logistic regression; class balance; feature combination; classification



**Citation:** Zhang, S.; Yuan, Y.; Yao, Z.; Yang, J.; Wang, X.; Tian, J. Coronary Artery Disease Detection Model Based on Class Balancing Methods and LightGBM Algorithm. *Electronics* **2022**, *11*, 1495. <https://doi.org/10.3390/electronics11091495>

Academic Editors: Stefanos Kollias and Jun-Ho Huh

Received: 21 March 2022

Accepted: 2 May 2022

Published: 6 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the development of economy and the aged tendency of population, the burden of disease and factors of death have changed dramatically all over the world. In the light of the analysis data of the World Health Statistics 2019, noncommunicable diseases have become the leading cause of death globally. In 2016, 41 million deaths, accounting for about 71% of global deaths, were caused by noncommunicable diseases. The main noncommunicable diseases incorporate cardiovascular disease, cancer, diabetes and chronic respiratory diseases [1]. Among them, cardiovascular disease (CVD) has become the leading cause of premature death and rising healthcare costs. In 2019, the worldwide cases of CVD

were 523 million, and the deaths caused by CVD reached 18.5 million, which accounted for approximately one-third of all deaths universally [2–4]. CVD mainly includes ischemic heart disease (IHD) and stroke. By 2019, there were 197 million IHD patients in the world. Additionally, the number of disability-adjusted life years (DALYs) because of IHD reached 182 million [5,6]. The detailed analysis of the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) 2019 shows that (See Table 1), in all age groups, the top three causes of DALYs are neonatal disorders, IHD and stroke, respectively. In the 50 to 74 years age group and 75 years and older age group, the top two causes of DALYs are IHD and stroke. Additionally, in the 25 to 49 years age group, the top three reasons of DALYs are road injuries, HIV/AIDS and IHD [6]. It's clear that IHD, a typical type of CVD, has become a major cause of death and disability.

**Table 1.** Top three causes of whole world DALYs and percentage of all DALYs in 2019.

Order	Causes <sup>1</sup>	Percentag <sup>1</sup>	Causes <sup>2</sup>	Percentag <sup>2</sup>	Causes <sup>3</sup>	Percentag <sup>3</sup>	Causes <sup>4</sup>	Percentag <sup>4</sup>
1	Neonatal disorders	7.3	Road injuries	5.1	IHD	11.8	IHD	16.2
2	IHD	7.2	HIV/AIDS	4.8	Stroke	9.3	Stroke	13.0
3	Stroke	5.7	IHD	4.7	Diabetes	5.1	COPD	8.5

<sup>1</sup> In all age groups. <sup>2</sup> In 25 to 49 years age group. <sup>3</sup> In 50 to 74 years age group. <sup>4</sup> In 75 years and older age group. DALY, disability-adjusted life-year. IHD, ischemic heart disease. COPD, chronic obstructive pulmonary disease. HIV, human immunodeficiency virus. AIDS, acquired immunodeficiency syndrome. The data in the table are from [6].

IHD, namely, coronary atherosclerotic heart disease (coronary artery disease (CAD) for short), refers to one kind of heart disease that manifests myocardial ischemia, hypoxia or necrosis caused by coronary artery narrowing or occlusion, which is caused by atherosclerosis. To be consistent with other literature, coronary artery disease (CAD) is also used in the later part of the article. According to the anatomical structure of coronary artery, there are three main arteries supplying blood to the heart, namely, (1) left anterior descending coronary artery (LAD), (2) left circumflex coronary artery (LCX) and (3) right coronary artery (RCA). CAD occurs when the lumen of any one of the three coronary arteries is narrowed by 50% or more [7].

CAD is a highly lethal disease, so early detection and diagnosis is crucial for saving lives and improving prognosis. At present, the auxiliary examinations used in CAD diagnosis in medical include laboratory examination, electrocardiogram (ECG) examination, radionuclide examination, multislice spiral CT coronary angiography imaging, echocardiography and coronary angiography. Other examination methods include fractional flow reserve (FFR), optical coherence tomography and intravenous ultrasound (IVUS). Among them, coronary angiography is regarded as the gold standard for the diagnosis of CAD. Unfortunately, many developing and low-income countries do not have the equipment and specialized doctors to perform coronary angiography. Moreover, imaging examinations with high cost and technical requirements are difficult to be popularized in many countries and regions. Therefore, more and more doctors, scholars and researchers are devoted to finding other methods that can detect and diagnose CAD early. Machine learning and data mining technology are such a field that they focus on. In the past few years, machine learning and data mining technology have been broadly applied in various aspects of medical researches, for instance disease screening, disease risk stratification, disease prediction and assistant decision-making [8–10].

Similarly, machine learning technology has also played an important role in the prediction tasks of CAD. Numerous studies have been carried out. According to the data types used in study, these studies can be divided into studies based on the ECG signals [11–22], studies based on the imaging data [23–29], and studies based on the data of multiple routine examination items. As mentioned above, many imaging examinations are difficult to popularize in some countries and regions. Therefore, there are certain regional

restrictions on the application of research based on the imaging data. The research based on the ECG signal that is easy to obtain can only identify the CAD with significant ECG changes. Nevertheless, for some hidden CAD, it is difficult to diagnose these CAD timely and accurately. The research based on the data of multiple routine examination items has the advantages of easy access to data and comprehensive reflection of CAD. Therefore, this study proposed a machine learning model for early and accurate detection of CAD based on conventional medical examination data. This study was carried out on Z-Alizadeh Sani dataset, which is the latest dataset with multiple examination indicators.

The rest of the paper is arranged as follows. Section 2, related work, summarizes the related research based on Z-Alizadeh Sani dataset in the past decades. Section 3 introduces the dataset used in our research. An introduction of our proposed method is exhibited in Section 4. Section 5 shows detailed information about the experiments and results, and provides a detailed analysis of the results. Section 6 discusses this study. The conclusion of the study is expressed in Section 7.

## 2. Related Work

In this section, we review the application and results of machine learning technology on CAD prediction tasks in the past ten years. We emphatically analyze the progress in technology and design of research based on the data of multiple routine examination items.

Alizadehsani et al. applied feature selection and feature construction techniques to process input data, and used multiple classification models on it. Finally, 94.08% classification accuracy was obtained on the sequence minimum optimization (SMO) model [30]. Elham et al. developed a technique called heterogeneous hybrid feature selection to select features that were related to CAD diagnosis from the Z-Alizadeh Sani dataset. Then, they applied two oversampling techniques, namely, synthetic minority oversampling technology (SMOTE) and adaptive synthetic (ADASYN) to deal with the imbalanced class problem in dataset. The classification accuracy gained was 92.58% [31]. Arabasadi et al. improved the classification performance of neural network on Z-Alizadeh Sani dataset by using a genetic algorithm to increase the initial weight value of the neural network. The classification performance with accuracy of 93.85%, specificity of 92% and sensitivity of 97% was achieved [32]. Alizadehsani et al. used the meta cost-sensitive algorithm to distinguish patients with CAD from healthy individuals. Naive bayes, support vector machines (SVM), k-nearest neighbors (KNN), SMO, and C4.5 algorithm were used for classification. The best classification result was obtained by SMO algorithm with accuracy of 92.09% and sensitivity of 97.22% [33]. Zomorodi-Moghadam et al. proposed a method discovering the rules of CAD classification. By using the feature selection method based on particle swarm optimization (PSO) and multi-objective evolutionary search, this method selected the features most related to CAD classification, and formed two different rule sets with 11 features and 13 features, respectively. The classification accuracy was evaluated on two rule sets. The results showed that this method has the ability to generate effective CAD detection rules [34].

Abdar et al. proposed a new nested integrated kernel support vector classification (NE-nu-SVC) model. The model improved the performance of traditional machine learning methods by applying ensemble learning technology, feature selection method and data balancing method. The model obtained 94.66% classification accuracy [35]. Abdar et al. introduced a method called N2Genetic optimizer to improve the performance of traditional algorithms. On the Z-Alizadeh Sani dataset, the method achieved an accuracy of 93.08% and F1 score of 91.51% [36]. Alizadehsani et al. realized the prediction of CAD by constructing a classifier for each coronary artery of LAD, LCX, and RCA. It was used to the extended Z-Alizadeh Sani dataset containing 500 patients and achieved 96.40% CAD detection accuracy [37]. Shahid et al. utilized four different feature selection means: Relief-F, Fisher, Weight by SVM, and Minimum Redundancy Maximum Relevance to improve the performance of emotional neural networks (EmNNs) on Z-Alizadeh Sani dataset. Finally, they obtained 88.34% classification accuracy [38]. Wang et al. developed a two-level

stacking model based on stacking ensemble learning idea for CAD detection. This model showed 95.43% detection accuracy on Z-Alizadeh Sani dataset [39]. Tama et al. designed a two-tier integration framework. The first layer was feature selection, which integrated two methods of feature selection, namely correlation-based feature selection (CFS) and PSO. The second layer was classifier modeling, which mixed the class label of three integrated learners through the stacking architecture. The integration framework achieved 98.13% accuracy on the Z-Alizadeh Sani dataset [40]. Gupta et al. proposed a system (C-CADZ) for CAD detection. C-CADZ automatically realized feature extraction, feature selection, class balance and model prediction. This system used random forest (RF) and extreme tree (ET) models as machine learning classifiers and obtained 97.37% prediction accuracy on Z-Alizadeh Sani dataset [41]. Kolukisa et al. processed the Z-Alizadeh Sani dataset by applying linear discriminant analysis (LDA) dimensionality reduction technique and a hybrid feature selection technique that combined four feature selection methods, namely, gain ratio (GR), information gain (IG), Relief-F (RF) and chi-square (CS) test. Finally, a multiclass fisher linear discriminant analysis (FLDA) ensemble classifier was used to classify, and the accuracy was 92.07% [42].

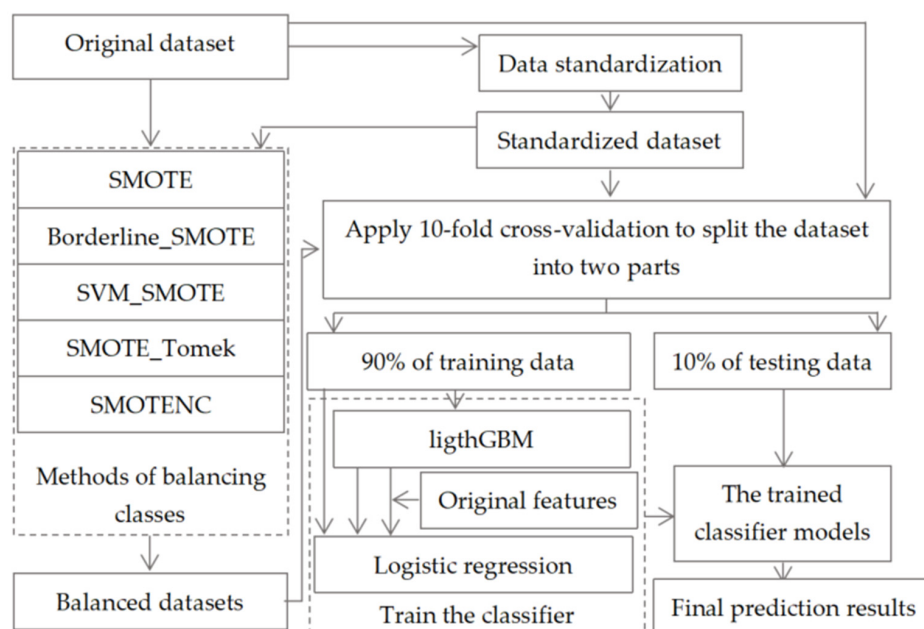
Dekamin et al. adopted K-means algorithm to preprocess data and used naive bayes, KNN and decision tree to classify. This method achieved 90.91% efficiency [43]. Alizadehsani et al. performed CAD prediction research based on laboratory and echocardiography data. This research used SMO, naive Bayes, C4.5 and AdaBoost algorithms for classification prediction and achieved a classification accuracy of more than 82% [44]. Yadav et al. proposed a CAD prediction method based on association rule mining, and obtained 92.09% classification accuracy on SMO classification algorithm [45]. Ghiasi et al. applied a decision tree learning algorithm called CART to the detection of CAD, and achieved 92.41% accuracy [46]. Joloudari et al. proposed a hybrid machine learning model. The model took SVM as the basic classifier, used analysis of variance as the kernel function of SVM, and used genetic optimization algorithm for feature selection. This model obtained 89.45% classification accuracy [47]. In our previous research, we applied four feature processing techniques and two kinds of class balancing methods to develop a CAD prediction model based on Random Forest algorithm and XGBoost algorithm. The model effectively realized the early detection of CAD and achieved 94.7% prediction accuracy [48].

In the past ten years, researchers have carried out a large number of exploratory research on CAD prediction based on machine learning technology on the dataset containing the results of multiple routine examination items. In reference [49], Alizadehsani et al. have analyzed 149 literatures related to the subject of machine learning-based CAD prediction published from 1992 to 2019. Reference [49] mainly summarized the investigated papers from the aspects of classification algorithm, feature selection algorithm and model evaluation. At the level of classification algorithm, most studies have applied traditional machine learning algorithms, such as SVM, Naive Bayes, KNN, decision tree, SMO, C4.5 and artificial neural network. Among them, artificial neural network and decision tree algorithm were the two most widely used algorithms. In the application of feature selection algorithm, the most commonly used methods include information gain, Gini index, PCA and weight by SVM. In the model evaluation, the application proportions of accuracy, recall, specificity and precision were 96%, 68.8%, 63.2% and 65.6%, respectively. The application proportions of F measurement and AUC was less than 20%.

Through the analysis, it can be found that the previous studies have made some achievements and progress, but there are still some limitations. For example, firstly, in the application of classification algorithm, there is less research on the application of classical logistic regression algorithm suitable for small datasets. Secondly, these kinds of data sets often have the problem of class imbalance. Several articles used some common sampling methods and cost sensitive algorithms to solve this problem, such as SMOTE algorithm and ADASYN algorithm. However, there are few studies on the application of more sampling methods. In addition, in terms of model evaluation indicators, the application of AUC and F measurement indicators is insufficient. However, these two indexes are very important

to evaluate the performance of the CAD prediction model. Especially in the dataset with class imbalance problem, the accuracy index often cannot accurately reflect the ability of the model.

In view of this, in this study, we proposed a machine learning model for early, accurate and rapid detection of CAD based on common medical detection data. The model used the classical logistic regression algorithm, which is the most commonly used in medical model research as the classifier. Additionally, we applied the advantages of feature selection and feature combination of tree models to solve the problem of manual feature engineering in logical regression algorithm. At the same time, five different sampling methods were applied to solve the class imbalance problem of Z-Alizadeh Sani dataset. Accuracy, recall, specificity, precision, F1 score, ROC and AUC were used to evaluate the performance of the model. Ten-fold cross-validation technology was also applied in the study. In addition, according to the characteristics of the dataset, we also adopted appropriate preprocessing methods. The frame diagram of our proposed machine learning model for CAD prediction is shown in Figure 1. The list of abbreviations used in this paper is shown in Table A1 of Appendix A.



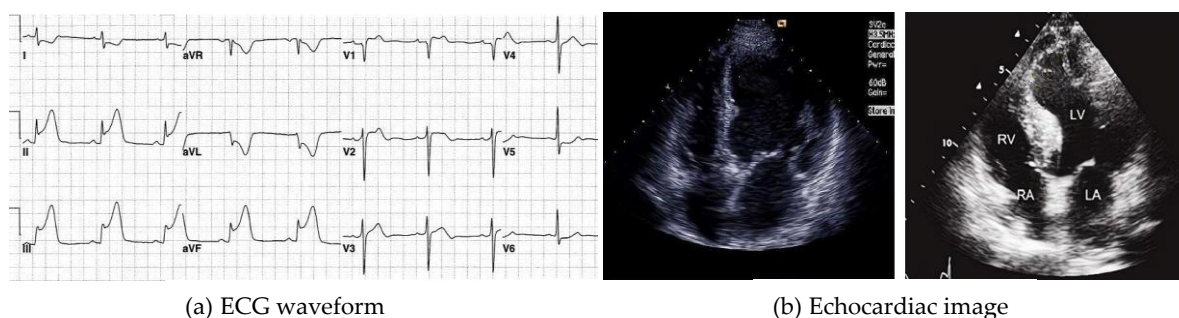
**Figure 1.** The frame diagram of our proposed machine learning model.

### 3. Dataset

From UCI Machine Learning Repository we downloaded the Z-Alizadeh Sani dataset used by us. The Z-Alizadeh Sani dataset is the latest dataset with multiple examination indicators. This dataset contains 303 medical records derived from 303 cases who visited Shaheed rajaei hospital because of chest pain. Each record includes 55 features belonging to four categories. Four categories are demographic features; symptoms and physical examination; ECG; and laboratory tests and echocardiography features. A record is a sample. These 303 samples belong to two classes, namely, CAD class and normal class. When the stenosis of coronary arteries lumen of a sample reaches or exceeds 50%, this sample is classified as CAD class; otherwise it belongs to the normal class. Accordingly, in 303 samples, 216 instances accounting for 71.29% are CAD class, and 87 instances accounting for 28.71% are normal class [30]. Details of the Z-Alizadeh Sani dataset are shown in Table 2. To intuitively explain the features of Z-Alizadeh Sani dataset, in Figure 2, we showed an ECG waveform and an echocardiatic image. The ECG features of Z-Alizadeh Sani dataset are obtained by professional doctors analyzing the ECG waveform in Figure 2a. Similarly, the echocardiography features of Z-Alizadeh Sani dataset are obtained by professional doctors detecting the echocardiatic image in Figure 2b.

**Table 2.** Detailed information of each feature and feature attribute in 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
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
Laboratory Tests and Echocardiography	BBB	N, LBBB, RBBB	categorical
	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



**Figure 2.** An example of an ECG waveform and an echocardiatic image (pictures from the Internet).

The 55 features of Z-Alizadeh Sani dataset can be split into two types: continuous features and categorical features. The statistical description of continuous features is shown in Table 3.

**Table 3.** Statistical description of continuous features of Z-Alizadeh Sani dataset.

Feature	Min	Max	Ave	Med	1th	5th	10th	50th	90th	95th	99th
Age	30	86	58.9	58.00	36.08	43.00	47.00	58.00	73.00	76.00	81.00
Weight	48	120	73.83	74.00	50.00	55.00	60.00	74.00	89.00	94.80	107.88
Length	140	188	164.72	165.00	145.00	150.00	152.00	165.00	176.60	179.00	185.96
BMI	18.12	40.90	27.25	26.78	18.83	21.10	22.31	26.78	33.21	34.88	38.24
BP	90	190	129.55	130.00	90.00	100.00	110.00	130.00	160.00	160.00	180.00
PR	50	110	75.14	70.00	60.00	64.00	70.00	70.00	87.20	90.00	109.60
FBS	62	400	119.18	98.00	69.04	77.00	80.00	98.00	193.20	223.80	358.40
CR	0.50	2.20	1.06	1.00	0.60	0.70	0.80	1.00	1.40	1.50	1.90
TG	37	1050	150.34	122.00	43.12	67.40	76.00	122.00	250.00	309.00	469.20
LDL	18	232	104.64	100.00	30.24	52.60	64.40	100.00	154.20	170.00	212.60
HDL	15.9	111.0	40.23	39.00	18.16	25.20	28.00	39.00	53.00	55.80	81.40
BUN	6	52	17.50	16.00	8.00	10.00	11.00	16.00	25.60	31.80	42.92
ESR	1	90	19.46	15.00	1.04	3.00	4.00	15.00	41.00	51.00	79.84
HB	8.9	17.6	13.15	13.20	9.00	10.02	11.00	13.20	15.10	15.68	17.19
K	3.0	6.6	4.23	4.20	3.20	3.50	3.70	4.20	4.80	5.00	5.40
Na	128	156	141.00	141.00	130.04	135.00	137.00	141.00	145.00	147.00	153.00
WBC	3700	18,000	7562.05	7100.00	3812.00	4700.00	5100.00	7100.00	10,700.00	12,100.00	16,960.00
Lymph	7	60	32.40	32.00	9.04	15.20	19.00	32.00	44.00	49.00	58.96
Neut	32	89	60.15	60.00	35.12	44.00	49.00	60.00	73.60	78.00	84.96
PLT	25	742	221.49	210.00	118.44	158.20	170.00	210.00	293.60	331.60	391.52
EF-TTE	15	60	47.23	50.00	20.00	26.00	35.00	50.00	55.00	55.00	60.00

Min, minimum value. Max, maximum value. Ave, average value. Med, median. th, th percentile.

By analyzing Tables 2 and 3, it can be seen that Z-Alizadeh Sani dataset has the following characteristics: (1) Exertional CP feature has the same values on 303 samples. In other words, External CP feature does not contribute to the classification prediction of CAD; (2) the dimensions of continuous features are not uniform, and the distribution of continuous features has a certain skewness; and (3) the dataset has class imbalance problem.

#### 4. Proposed Method

This section describes the machine learning techniques used by the study in detail.

##### 4.1. Preprocessing of Data

In view of the characteristics of Z-Alizadeh Sani dataset, we processed the dataset as follows: (1) delete the External CP feature directly; and (2) standardize the continuous features.

The method of data standardization is as follows:

$$\mathbf{X} = \left\{ \chi^{(1)}, \chi^{(2)}, \chi^{(3)} \dots, \chi^{(m)} \right\} \quad (1)$$

where  $\mathbf{X}$  is sample set,  $\chi^{(1)}, \chi^{(2)}, \chi^{(3)} \dots, \chi^{(m)}$  are samples of  $\mathbf{X}$ ,  $m$  is the number of samples. Suppose that sample  $\chi^{(i)}$  of  $\mathbf{X}$  has  $n$  features, that is:

$$\chi^{(i)} = \{\chi_1^{(i)}, \chi_2^{(i)}, \chi_3^{(i)} \dots, \chi_n^{(i)}\} \tag{2}$$

Then, for the continuous feature  $j$  there is:

$$\mu_j = \frac{1}{m} \sum_{i=1}^m \chi_j^i \tag{3}$$

where  $\chi_j^{(i)}$  is the value of feature  $j$  on sample  $\chi^{(i)}$ .  $\mu_j$  is the average value of feature  $j$ . Additionally, there is:

$$\sigma_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (\chi_j^i - \mu_j)^2} \tag{4}$$

Thereinto,  $\sigma_j$  is the standard deviation of feature  $j$ . Then:

$$\chi_{j*}^i = \frac{\chi_j^i - \mu_j}{\sigma_j} \tag{5}$$

$\chi_{j*}^i$  is the value of feature  $j$  on sample  $\chi^{(i)}$  after standardization.

#### 4.2. Methods of Balancing Classes

We applied five different methods of classes balancing on Z-Alizadeh Sani dataset. We agree that the sample set of minority class is  $S_{\min}$ , the sample set of majority class is  $S_{\max}$ ,  $x^1, x^2, x^3, \dots, x^m$  are the samples of  $S_{\min}$ ,  $m$  is the number of samples of  $S_{\min}$ , and  $x^i$  refers to any sample of  $S_{\min}$ .

##### 4.2.1. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) [50–54] is one of the most commonly used and classic oversampling methods. Simply put, SMOTE synthesizes new samples artificially by analyzing the distance between minority samples and their nearest neighbors. The specific process of synthesizing samples is as follows: (1) distance calculation: calculate the distance between  $x^i$  and all samples of  $S_{\min}$ ; thus, the  $k$ -nearest neighbors of sample  $x^i$  are obtained; (2) nearest neighbors selection: set the number of oversampling of each sample of minority class according to the ratio of  $S_{\min}$  and  $S_{\max}$ , assuming this number is  $n$ . Then,  $n$  samples are randomly selected from the  $k$ -nearest neighbors of sample  $x^i$ . Here, we assume that  $x^j$  is one of the  $n$ -nearest neighbors samples; (3) new samples synthesis: the new sample  $x^{new}$  between the sample  $x^i$  and  $x^j$  is synthesized according to the following formula:

$$x^{new} = x^i + rand(0, 1) \times |x^i - x^j| \tag{6}$$

where  $rand(0, 1)$  refers to generate a random number uniformly distributed between 0 and 1,  $|x^i - x^j|$  is the distance between the sample  $x^i$  and  $x^j$ .

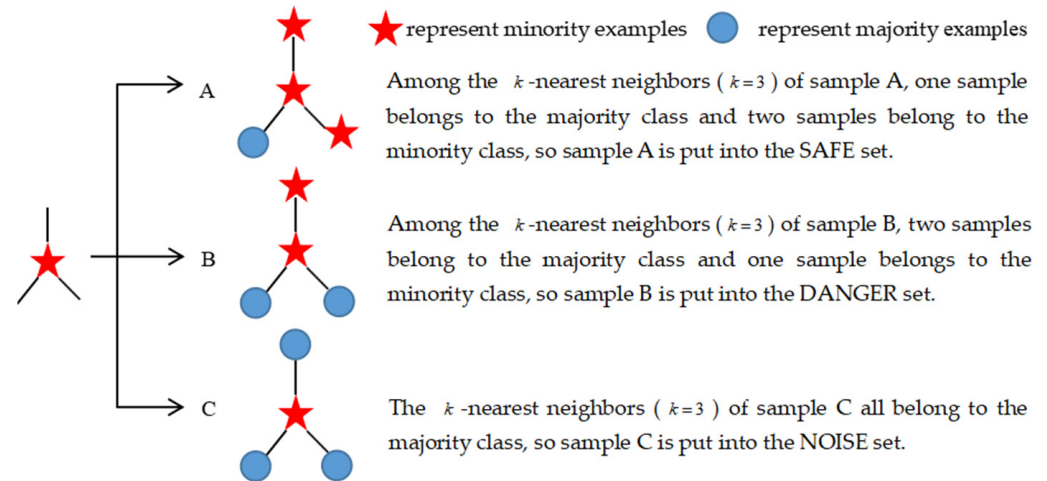
All new samples are synthesized in the same way until the class balance requirements are met. The essence of SMOTE algorithm is to select a random point on the connection line of two specific samples as a new sample. This method increases the number of minority samples effectively.

##### 4.2.2. Borderline\_SMOTE

Unlike SMOTE, which synthesizes new samples for each minority sample, Borderline\_SMOTE algorithm [55] only resamples or strengthens the minority examples at the



class borderline. Firstly, the  $k$ -nearest neighbors of sample  $x^i$  are found. Then, the number of samples belonging to the majority class in the  $k$ -nearest neighbors of sample  $x^i$  is analyzed, assuming this number is  $b$ . When  $b < \frac{k}{2}$ , sample  $x^i$  is put into the SAFE set. When  $b > \frac{k}{2}$ , sample  $x^i$  is put into the DANGER set. When  $b = k$ , sample  $x^i$  is put into the NOISE set, as shown in Figure 3. Finally, the minority samples belonging to the DANGER set are resampled to synthesize new samples. The process of new sample synthesis is the same as SMOTE.



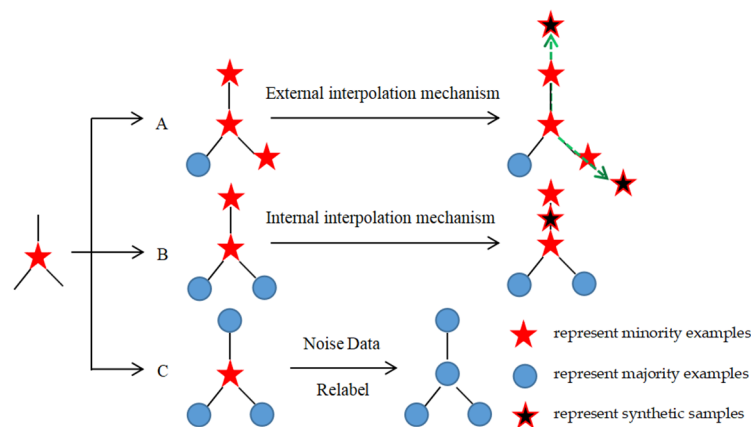
**Figure 3.** Standard for Borderline\_SMOTE algorithm to assign minority samples to SAFE set, DANGER set and NOISE set.

Borderline\_SMOTE algorithm includes Borderline\_SMOTE1 and Borderline\_SMOTE2. Borderline\_SMOTE1 will randomly select minority sample from the  $k$ -nearest neighbors of sample  $x^i$  to synthesize new samples. While Borderline\_SMOTE2 will randomly select any sample from the  $k$ -nearest neighbors of sample  $x^i$  to synthesize new samples. In this article, we applied Borderline\_SMOTE1.

#### 4.2.3. SVM\_SMOTE

SVM\_SMOTE algorithm [56] is the combination of SVM algorithm and SMOTE algorithm. Similar to Borderline\_SMOTE algorithm, SVM\_SMOTE algorithm also focuses on the minority samples at the class borderline and only resamples or strengthens these samples. Firstly, the SVM classifier is applied on the training set to obtain the support vector, which is approximately the boundary region. Then, the new minority samples will be synthesized according to the following decision mechanism. The decision mechanism for synthesizing minority class samples depends on the distribution density of majority class samples around minority class support vector samples (See Figure 4). When more than half of the  $k$ -nearest neighbor samples of a minority class support vector sample belong to the minority class, the new minority class samples will be synthesized by an external interpolation mechanism. When more than half of the  $k$ -nearest neighbor samples of a minority class support vector sample belong to the majority class, the new minority class samples will be synthesized by an internal interpolation mechanism. When all the  $k$ -nearest neighbor samples of a minority class support vector sample belong to the majority class, it is considered that the minority class support vector sample is noise data and should be relabeled.

The essence of SVM\_SMOTE algorithm is over-sampling based on support vector. Near the class boundary approximated by the support vector, different decision mechanisms are selected to synthesize the minority class samples according to the distribution density of the majority class samples around the minority class support vector. After applying SVM\_SMOTE algorithm, the minority groups can be expanded to areas with low sample density of the majority class.



**Figure 4.** The decision mechanism of SVM\_SMOTE algorithm for synthesizing new minority samples. A, B and C are minority class support vector samples.

#### 4.2.4. SMOTE\_Tomek

The above three methods are all oversampling methods, which achieve class balance by increasing the number of minority class samples. However, the SMOTE\_Tomek algorithm [57] introduced in this section is a comprehensive sampling method combining oversampling and undersampling. Before introducing the SMOTE\_Tomek algorithm in depth, we need to understand the definition of Tomek links. Simply put, Tomek links is defined as a pair of connections between two nearest neighbor samples belonging to the opposite class. Assuming that  $x_i$  and  $x_j$  are two samples belonging to different class,  $d(x_i, x_j)$  is the distance between  $x_i$  and  $x_j$ . If there is no sample  $x_e$ , which can make  $d(x_e, x_i) < d(x_i, x_j)$  or  $d(x_e, x_j) < d(x_i, x_j)$ , the sample pair  $(x_i, x_j)$  is a Tomek link. Two samples of a Tomek link, either one of them is noise, or both of them are in the boundary area. Therefore, Tomek links can be used for both data undersampling and data cleaning. When Tomek links are used as an undersampling method, samples belonging to the majority class in Tomek links are deleted. When Tomek links are used as a data cleaning method, two samples in Tomek links are deleted.

The SMOTE\_Tomek algorithm applied in this paper takes advantage of the data cleaning effect of Tomek links. Firstly, SMOTE algorithm was applied to the original dataset, and a balanced dataset  $T_{SMOTE}$  was obtained. Then, we found the Tomek links from  $T_{SMOTE}$  and deleted the two samples in Tomek links. Finally, a new balanced dataset was obtained for training and testing.

#### 4.2.5. SMOTENC

SMOTENC algorithm [51] is an oversampling method that can deal with categorical features. It is assumed that sample  $x_i$  is a minority class sample containing both continuous and categorical features, sample  $x_{new}$  is a new sample synthesized from sample  $x_i$  by SMOTENC algorithm, and feature  $j$  is one of the categorical features. The value of feature  $j$  of sample  $x_{new}$  is the value with the highest frequency among the values of feature  $j$  of the  $k$ -nearest neighbors of sample  $x_i$ .

#### 4.3. Feature Combination

In this study, logistic regression (lr) algorithm is used as a classifier for classification and prediction. Logistic regression is a classification model with the advantages of easy understanding, easy implementation and fast operation. However, when used for classification, a large number of feature engineering, such as feature extraction and feature combination, need to be carried out manually to improve the performance of lr classifier. Undoubtedly, it takes a lot of time to carry out feature engineering manually. Therefore, many researchers tried to improve the classification ability of lr through automatic feature engineering.

For a long time, decision tree model and neural network model have shown better performance in disease classification and prediction tasks [58–61]. The former has good interpretability, while the latter shows high efficiency on large data sets. The splitting process of the model based on decision tree algorithm in each node is the process of feature screening. Each leaf node on the decision tree is the output of the decision tree trained on a set of effective feature subsets. That is, in the decision tree, the path from the root node to each leaf node is a feature combination process. Each leaf node contains a set of unique feature combination information. Therefore, models based on decision tree algorithm often have good functions of feature selection and feature combination. Deepfm model is a classic model in neural network, which was proposed by Huawei Noah Ark laboratory in 2017 [62]. Deepfm model is the combination of factorization machines (FM) and neural network structure. Deepfm model can automatically learn low-order explicit feature combination and high-order implicit feature combination at the same time. Therefore, the Deepfm model can also be used for automatic feature combination.

Using the feature combination function of models based on decision tree algorithm and Deepfm model to improve the performance of lr classifier has been widely used in the prediction task of advertising click through rate, and has shown good performance. However, in the prediction task of CAD, the research on this method has not been carried out. Therefore, in this paper, we will utilize the feature combination advantages of decision tree model or Deepfm model to improve the performance of lr classifier on Z-Alizadeh Sani dataset.

To select the best tree model for feature combination, we applied five classification models based on ensemble learning technology to the original dataset. The five classification models are random forest (RF), extratrees, adaptive boosting (AdaBoost), eXtreme gradient boosting (XGBoost) and light gradient boosting machine (lightGBM). The performance results of the five models on the original dataset are shown in Table 4. It can be seen from Table 4 that on the original dataset the lightGBM model has the best performance. In addition, we also applied the DeepFM model to the original dataset. Unfortunately, the highest accuracy obtained by DeepFM model was only 83.52%. This may be related to the small sample size of the dataset. Therefore, we select the lighthGBM model with the best performance to play the function of feature combination for lr classifier.

**Table 4.** The performance results obtained by the five models on the original dataset.

Algorithms	Accuracy	Recall	F1	Precision	Specificity	AUC
RF	0.887 ± 0.056	0.909 ± 0.051	0.923 ± 0.038	0.940 ± 0.051	0.834 ± 0.117	0.92 ± 0.05
Extratrees	0.891 ± 0.045	0.932 ± 0.045	0.923 ± 0.033	0.916 ± 0.055	0.788 ± 0.123	0.91 ± 0.06
AdaBoost	0.904 ± 0.059	0.929 ± 0.067	0.934 ± 0.038	0.944 ± 0.050	0.842 ± 0.111	0.93 ± 0.05
XGBoost	0.894 ± 0.068	0.909 ± 0.072	0.929 ± 0.044	0.954 ± 0.046	0.856 ± 0.106	0.93 ± 0.05
lightGBM	0.911 ± 0.060	0.922 ± 0.068	0.940 ± 0.038	0.963 ± 0.041	0.881 ± 0.095	0.93 ± 0.05

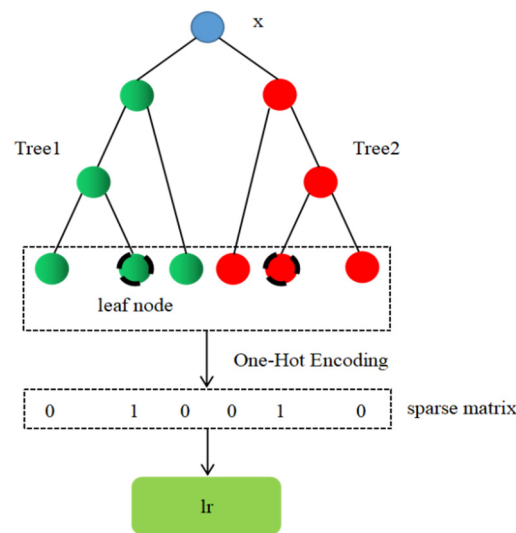
LightGBM [63] is a machine learning algorithm framework based on boosting ensemble learning technology proposed by Microsoft in 2017. Compared with GBDT (Gradient Boosting Decision Tree) algorithm and XGBoost algorithm, lightGBM algorithm can not only achieve the same prediction performance, but also has more obvious advantages in training speed and memory consumption. On the basis of GBDT algorithm and XGBoost algorithm, lightGBM algorithm integrates several optimization strategies, such as histogram algorithm, gradient-based one-side sampling (GOSS) algorithm, exclusive feature bundling (EFB) strategy, leaf-wise strategy, supporting category feature strategy and supporting efficient parallel strategy. These optimization strategies make lightGBM become a classification model with high efficiency, low consumption, more accuracy and more convenient. The parameter `boosting_type` was set to GBDT (Gradient Boosting Decision Tree) [64–67] algorithm when we used the lightGBM algorithm.

The method where lightGBM algorithm plays the function of feature combination for lr classifier is to use the output of lightGBM algorithm as the input of lr algorithm for training to obtain the final prediction output. This is the application of stacking ensemble

learning technology. It is worth noting that the input of lr algorithm is not the category labels (such as 0 or 1) predicted by lightGBM algorithm for each instance, but the index of leaf node of each instance on each decision tree. These indexes need to be processed by One-Hot Encoding before inputting lr algorithm. The specific process is as follows:

- (1) LightGBM model is applied to the training set, and the trained LightGBM classifier is obtained.
- (2) After the training, the index of leaf node on each decision tree in the lightGBM model is output for each instance of the training set. After all iterations, all indexes of each instance form a new set of features. At this time, a set of  $m \times n$ -dimensional dataset is formed, where  $m$  is the size of the training set, and  $n$  is the number of weak estimators (decision trees) in the lightGBM model.
- (3) The  $m \times n$ -dimensional dataset obtained in step (2) is encoded by One-Hot Encoding, and a sparse matrix  $M_{train}$  with  $m \times n \times l$ -dimensional is obtained, where  $l$  is the number of leaf nodes per decision tree in the lightGBM model. The sparse matrix  $M_{train}$  is the training set of lr algorithm.
- (4) lr model is applied to the sparse matrix  $M_{train}$ , and the trained lr classifier is obtained.
- (5) Similarly, the testing set is processed by steps (1)–(3) to obtain the sparse matrix  $M_{test}$ , and the sparse matrix  $M_{test}$  is entered into the trained lr classifier to obtain the final prediction.

The above is the process of realizing automatic feature combination based on lightGBM model (see Figure 5). In addition, we also tried another feature combination method, that is, the method of combination the sparse matrix output by ightGBM model with the original feature set. In other words, in this study, we tried two methods of feature combination. One is to input the sparse matrix  $M_{train}$  combined by lightGBM mode directly into the lr model for training. The other is to recombine the sparse matrix  $M_{train}$  combined by lightGBM mode with the training set of the original feature set and then input them into the lr model for training. For the sake of distinction, we record the lr classifier trained by the former as lightGBM + lr, and the lr classifier trained by the latter as lightGBM + LR.



**Figure 5.** The process of realizing automatic feature combination based on lightGBM model.

#### 4.4. Classification Algorithm

Logistic Regression (lr) algorithm [68] is one of the most commonly used exploration tools in medical research, especially in the medical binary prediction task, lr model has a wide range of applications. The classification decision-making process of lr algorithm for a specific sample is shown in Equations (7)–(10).

For binary classification problem, suppose any sample  $\mathbf{X}$  has:

$$\mathbf{X} = \{x_0, x_1, x_2, x_3, \dots, x_n\} \tag{7}$$

where  $n$  is the number of features contained in sample  $\mathbf{X}$ ,  $x_0$  is the bias term,  $x_i$  is the value of sample  $\mathbf{X}$  on the  $i$ th feature,  $i = 1, 2, 3, \dots, n$ .

Then, the probability  $\hat{P}$  (or called it as decision function  $h_{\theta}(\mathbf{X})$ ) that the lr model predicts sample  $\mathbf{X}$  as a positive class is calculated according to the following equation:

$$\hat{P} = h_{\theta}(\mathbf{X}) = \sigma(\mathbf{X}^T \boldsymbol{\theta}) \quad (8)$$

where  $\boldsymbol{\theta}$  is the weight vector and  $\boldsymbol{\theta} = \{\theta_0, \theta_1, \theta_2, \theta_3, \dots, \theta_n\}$ ,  $\theta_i$  is the weight value corresponding to  $x_i$ ,  $i = 0, 1, 2, 3, \dots, n$ .  $\sigma$  is the sigmoid function i.e.,  $\sigma(t) = \frac{1}{1 + \exp(-t)}$ , then:

$$\hat{P} = h_{\theta}(\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}^T \boldsymbol{\theta})} \quad (9)$$

$h_{\theta}(\mathbf{X})$  is the probability that the sample  $\mathbf{X}$  belongs to the positive class calculated by the lr model. Combined with the classification threshold, the prediction class of sample  $\mathbf{X}$  can be obtained. Assuming that the classification threshold is set to 0.5, the prediction function of lr model is:

$$\hat{y}(\mathbf{X}) = \begin{cases} 0, & \text{if } h_{\theta}(\mathbf{X}) < 0.5 \\ 1, & \text{if } h_{\theta}(\mathbf{X}) \geq 0.5 \end{cases} \quad (10)$$

where  $\hat{y}(\mathbf{X})$  represents the prediction function, 0 and 1 represent the class codes of negative class and positive class, respectively, 0.5 is the classification threshold.

The above is the process of lr model classifying a specific sample. It can be summarized as follows: firstly, the probability that sample  $\mathbf{X}$  in Equation (7) belongs to positive class is calculated through Equations (8) and (9). Then, the probability is compared with the classification threshold according to Equation (10). At this time, the prediction class of sample  $\mathbf{X}$  can be obtained.

In the training process of the model, we directly take the classification evaluation index as the objective function of key hyperparametric optimization. In order to take into account the performance of the two classes at the same time, we take the evaluation index F1 score as the objective function and use the learning curve to find the optimal hyperparameter. The hyperparameters setting of the proposed model and an example is shown in Table A2 of Appendix A.

## 5. Experiments Results

### 5.1. Evaluation Metrics

In order to comprehensively evaluate the classification performance and effectiveness of our proposed method, we applied accuracy, recall, F1 score, precision, specificity, ROC and AUC evaluation metrics. For the sake of expression of the significance and calculation formula of these evaluation metrics, we introduced the confusion matrix (See Table 5) first. The confusion matrix is a specific matrix used to visually present the performance of the algorithm. The confusion matrix of binary classification consists of two rows and two columns. Rows represent the true labels of the two classes in the dataset (denoted  $y_{\text{true}}$ ). Columns represent the predicted label of the two classes acquired by the model (denoted as  $y_{\text{pre}}$ ). As shown in Table 5, the confusion matrix of binary classification includes four indicators: TN, FN, FP and TP. The four indicators are defined as follows. We specified that the label of positive class is 1 and the label of negative class is 0.

**Table 5.** Confusion matrix.

	$y_{\text{pre}} = 0$	$y_{\text{pre}} = 1$
$y_{\text{true}} = 0$	TN	FP
$y_{\text{true}} = 1$	FN	TP

TN (true negative) refers to the number of correctly predicted samples in the samples with the real class label of 0.

FN (false negative) refers to the number of incorrectly predicted samples in the samples with the real class label of 1.

FP (false positive) refers to the number of incorrectly predicted samples in the samples the real class label of 0.

TP (true positive) refers to the number of correctly predicted samples in the samples with the real class label of 1.

#### 5.1.1. Accuracy

Accuracy refers to the proportion of samples that can be correctly predicted by the model in all samples. The calculation equation of accuracy is as follows. TN, TP, FN and FP refer to true negative, true positive, false negative and false positive, respectively.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (11)$$

Accuracy is one of the most frequently used and most important model performance evaluation metrics. However, in the dataset with a class imbalance problem, due to the influence of majority class samples, the accuracy is often difficult to accurately measure the classification ability of the model. Therefore, in the dataset with a class imbalance problem, in addition to accuracy, more evaluation indicators need to be applied.

#### 5.1.2. Recall

Recall refers to the proportion of samples that can be correctly predicted by the model in all samples with positive real class labels. Recall is an important indicator to measure the ability of model to identify positive samples. In medical models, it is necessary to pay attention to recall. Recall is calculated according to the following equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

where TP and FN are true positive and false negative, respectively. In medical application, the cost of undiagnosed positive cases and wrongly diagnosed negative cases is different. The former may cause loss of life, while the latter may lead to excessive treatment. Compared with the former, the latter costs less. At the time of diagnosis, doctors and patients pay more attention to the detection of positive cases. Therefore, the recall is one of the important indicators to judge whether the model can be applied in practice.

#### 5.1.3. Precision

Precision, like recall, is an important indicator to measure the ability of the model to correctly predict positive samples. Precision refers to the proportion of samples with positive real class labels among all samples predicted as positive by the model. According to the definition of precision, its calculation formula is as follows. TP and FP in the formula are true positive and false positive, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

#### 5.1.4. F1 Score

Sometimes the performance of the model evaluated by recall and precision may be show the opposite result, that is, one index has a good result but the other index has a poor result, so the ability of the model cannot be evaluated accurately. Therefore, F1 score is introduced. F1 score combines the results of recall and precision, and is the weighted harmonic mean of recall and precision. Only when the results of recall and precision

are good, the F1 score will be higher. The higher the F1 score is, the better the model classification effect is. The following is the calculation formula of F1 score.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

### 5.1.5. Specificity

Specificity refers to the proportion of samples that can be correctly predicted by the model in all samples with negative real class labels. Specificity measures the ability of the model to recognize negative samples. Specificity is calculated according to the following formula. TN and FP correspond to true negative and false positive, respectively.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (15)$$

### 5.1.6. ROC and AUC

Area under curve (AUC) is the area under the receiver operating characteristic (ROC) curve. The ROC curve is drawn with the false positive rate (FPR) as x-axis and the true positive rate (TPR) as y-axis. ROC curve intuitively reflects the relationship between specificity and recall. The value of AUC is between 0 and 1, when the value of x-axis (i.e., the false positive rate (FPR) of the model) is closer to 0, and the value of y-axis (i.e., the true positive rate (TPR) of the model) is closer to 1, the value of AUC is closer to 1. The closer the AUC value is to 1, the higher the prediction performance of the classifier.

## 5.2. Experimental Results

In this section, we reported the classification results obtained by our proposed model on various data sets used in our study. The classification results are described from the accuracy, recall, precision, F1 score, specificity, ROC and AUC indicators. In order to explore the impact of data preprocessing method and class balancing methods on the classification performance of the model, we applied five class balancing methods on the original dataset and the standardized dataset, respectively. After the above processing, a total of 12 datasets were generated for research. At the same time, in order to verify whether our proposed method has competitive advantage in CAD prediction task, we also output the prediction results of lightGBM model and lr model without feature combination on each dataset during the experiment. Therefore, we obtained four groups of experimental results on each dataset. To facilitate narration, we record the classifiers that produce the four groups of experimental results as lightGBM, lr, lightGBM + lr and lightGBM + LR, respectively.

### 5.2.1. Results Obtained on Source Dataset

The performance results of the classification models used for CAD prediction on the source dataset are reported in this section. There is a certain class imbalance in the source dataset, which shows that 216 instances belong to CAD class (accounting for 71.29%) and 87 instances belong to the normal class (accounting for 28.71%). The average results of 10 tests in terms of accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on the source dataset are shown in Table 6. Figure 6 is an intuitive display of Table 6. Figure 7 shows the ROC curve of the classification model on the 10 fold test set and the AUC value corresponding to each fold ROC curve.

**Table 6.** The average results of 10 tests obtained by classification models on the source dataset.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.911 ± 0.060	0.922 ± 0.068	0.940 ± 0.038	0.963 ± 0.041	0.881 ± 0.095	0.93 ± 0.05
lr	0.884 ± 0.052	0.911 ± 0.059	0.921 ± 0.032	0.935 ± 0.042	0.819 ± 0.099	0.93 ± 0.05
lightGBM + lr	0.907 ± 0.047	0.921 ± 0.059	0.937 ± 0.028	0.958 ± 0.038	0.873 ± 0.081	0.93 ± 0.06
lightGBM + LR	0.914 ± 0.045	0.925 ± 0.059	0.942 ± 0.029	0.963 ± 0.035	0.886 ± 0.074	0.93 ± 0.05

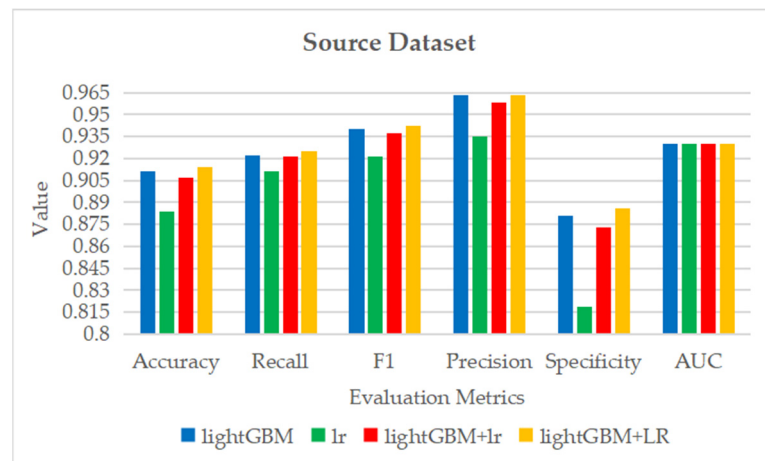


Figure 6. The histogram of Table 6.

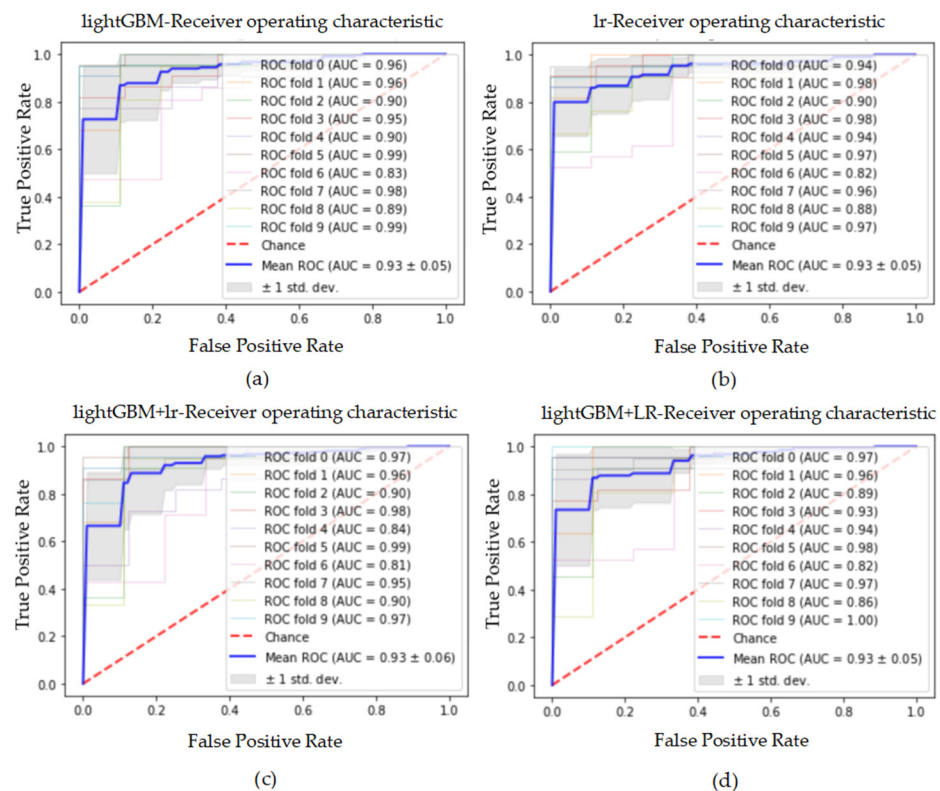


Figure 7. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the original dataset. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

It can be seen from Table 6 and Figure 6 that on the original dataset, the performance of lr model without feature combination processing is weaker than that of the ensemble algorithm lighGBM. However, the classification performance of lr model is significantly improved after combining our proposed feature combination processing method. On the original dataset with new features combined by lighGBM, lr model obtains the classification result better than lighGBM model, which is also the best classification result on the dataset. The highest accuracy is 91.4%, recall is 92.5%, F1 score is 94.2%, precision is 96.3%, specificity is 88.6%, and AUC is 0.93. Figure 7 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the original dataset. It can be seen from the curve that, due to the distribution difference of



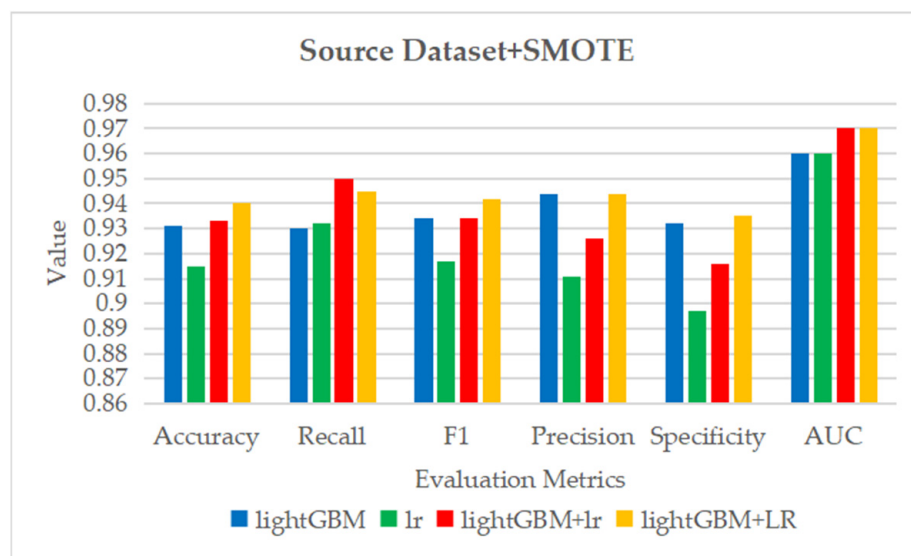
10 fold data, the AUC values obtained by the model on each fold test set are different. In addition, although the average AUC of the four classification models on 10 fold data are all about 0.93, the AUC disturbance of lightGBM + lr model, which has a standard deviation of 0.06, is slightly larger than that of the other three models.

### 5.2.2. Results Obtained on Dataset Processed by SMOTE

This part reports the performance results of the classification models for CAD prediction on the balanced dataset processed by SMOTE algorithm. The balanced dataset after SMOTE algorithm processing contains 432 sample instances, with CAD class and normal class accounting for 50%, respectively. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on the dataset processed by SMOTE are shown in Table 7. Figure 8 is a visual display of Table 7. Figure 9 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by SMOTE and the AUC value corresponding to each fold ROC curve.

**Table 7.** The average results of 10 tests obtained by classification models on the dataset processed by SMOTE.

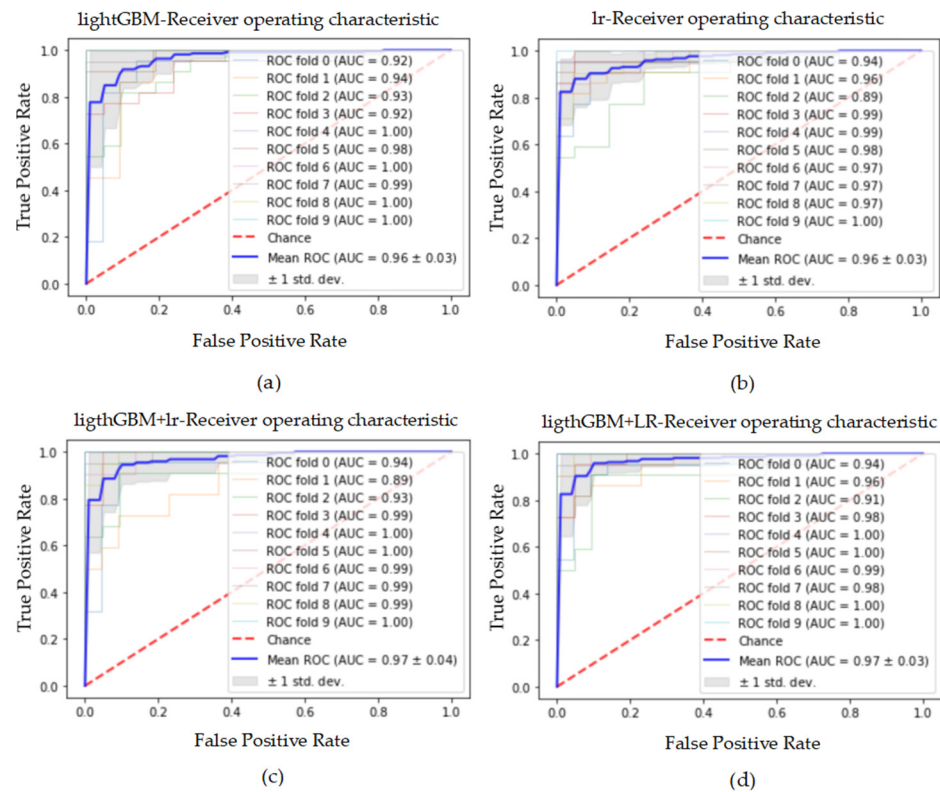
Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.931 ± 0.055	0.930 ± 0.090	0.934 ± 0.048	0.944 ± 0.035	0.932 ± 0.039	0.96 ± 0.03
lr	0.915 ± 0.057	0.932 ± 0.096	0.917 ± 0.049	0.911 ± 0.058	0.897 ± 0.047	0.96 ± 0.03
lightGBM + lr	0.933 ± 0.054	0.950 ± 0.083	0.934 ± 0.047	0.926 ± 0.052	0.916 ± 0.053	0.97 ± 0.04
lightGBM + LR	0.940 ± 0.048	0.945 ± 0.078	0.942 ± 0.043	0.944 ± 0.041	0.935 ± 0.037	0.97 ± 0.03



**Figure 8.** The histogram of Table 7.

As can be seen from Table 7 and Figure 8, compared with Table 6 and Figure 6, the classification performance of the four classification models in terms of accuracy, recall, specificity and AUC has been significantly improved on the balanced dataset processed by SMOTE. The performance of four classification models on the dataset used in this section is also different. Similarly, on the balanced dataset after SMOTE processing, the performance of lr model without feature combination processing is still worse than lightGBM. After the feature combination processing of lightGBM model, the performance of lr classifier in accuracy, recall, F1 score, precision, specificity and AUC has been significantly improved. From the point of view of all classification indicators, the best classification performance appears in the lightGBM + LR model, with accuracy of 94.0%, recall of 94.5%, F1 score 94.2%, precision of 94.4%, specificity of 93.5% and AUC of 0.97. However, from the recall alone, the best recall result appears in the lightGBM + lr model, which is 95.0%. Figure 9

shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset balanced by SMOTE. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. However, it is obvious that model lightGBM + LR has the best AUC value.



**Figure 9.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by SMOTE. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

### 5.2.3. Results Obtained on Dataset Processed by Borderline\_SMOTE

In this part, we describe the classification results of the models on the balanced dataset processed by Borderline\_SMOTE. The results are described based on six classification indicators: accuracy, recall, precision, F1 score, specificity and AUC. The balanced dataset processed by Borderline\_SMOTE contains 432 sample instances, including 216 cases of CAD class and 216 cases of normal class. The average results of 10 tests on six classification indicators obtained by classification models on the balanced dataset processed by Borderline\_SMOTE are exhibited in Table 8. Figure 10 is an intuitive display of Table 8. Figure 11 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by Borderline\_SMOTE and the AUC value corresponding to each fold ROC curve.

**Table 8.** The average results of 10 tests obtained by classification models on the dataset processed by Borderline\_SMOTE.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.933 ± 0.041	0.938 ± 0.081	0.935 ± 0.035	0.939 ± 0.042	0.928 ± 0.033	0.97 ± 0.03
lr	0.903 ± 0.052	0.919 ± 0.087	0.904 ± 0.046	0.897 ± 0.059	0.887 ± 0.049	0.96 ± 0.03
lightGBM + lr	0.933 ± 0.047	0.949 ± 0.068	0.933 ± 0.046	0.921 ± 0.057	0.917 ± 0.056	0.97 ± 0.03
lightGBM + LR	0.938 ± 0.053	0.938 ± 0.081	0.940 ± 0.048	0.948 ± 0.058	0.938 ± 0.058	0.97 ± 0.03

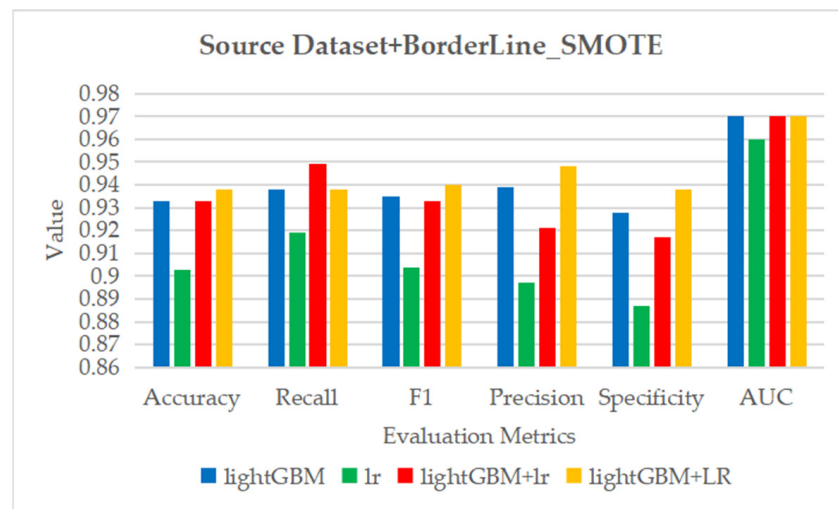


Figure 10. The histogram of Table 8.

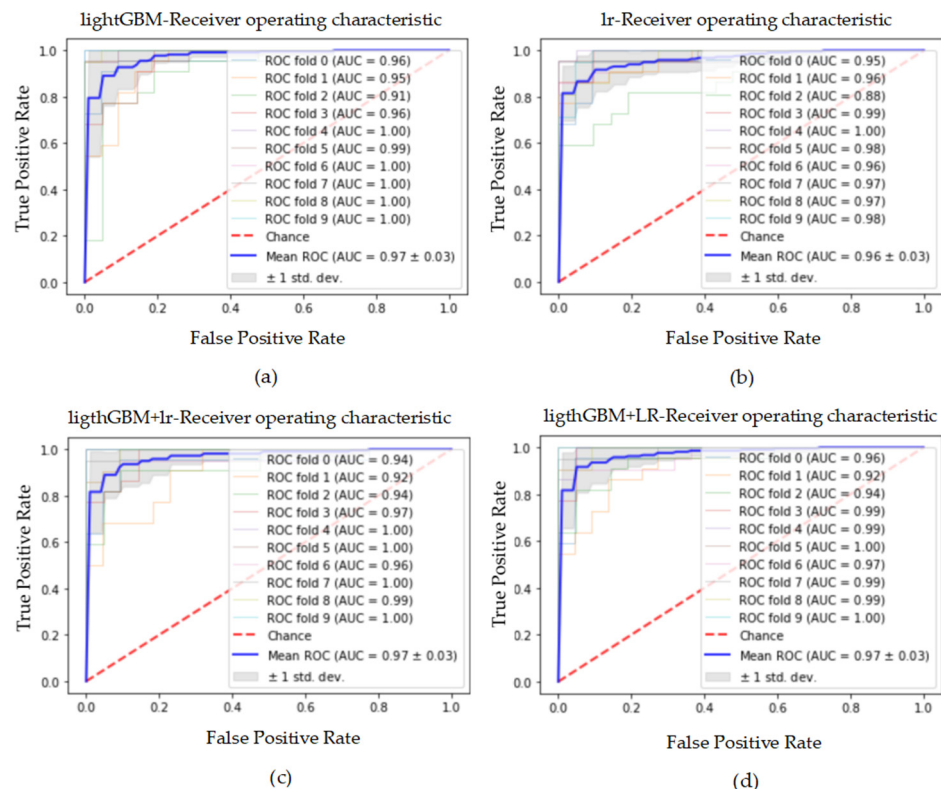


Figure 11. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by Borderline\_SMOTE. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

By comparing Tables 6 and 8, we can find that Borderline\_SMOTE class balancing method improves the classification performance of models in terms of accuracy, recall, specificity and AUC. On this dataset, the feature combination function of lightGBM model can still significantly improve the performance of lr model in terms of accuracy, recall, F1 score, precision, specificity and AUC. The best classification results in this dataset are obtained by lr model on the dataset combining the combined features output by lightGBM and the original feature set, with accuracy of 93.8%, recall of 93.8%, F1 score 94.0%, precision of 94.8%, specificity of 93.8% and AUC of 0.97. The best recall result in this section appears in the lightGBM + lr model, which is 94.9%. Similarly, when comparing the classification

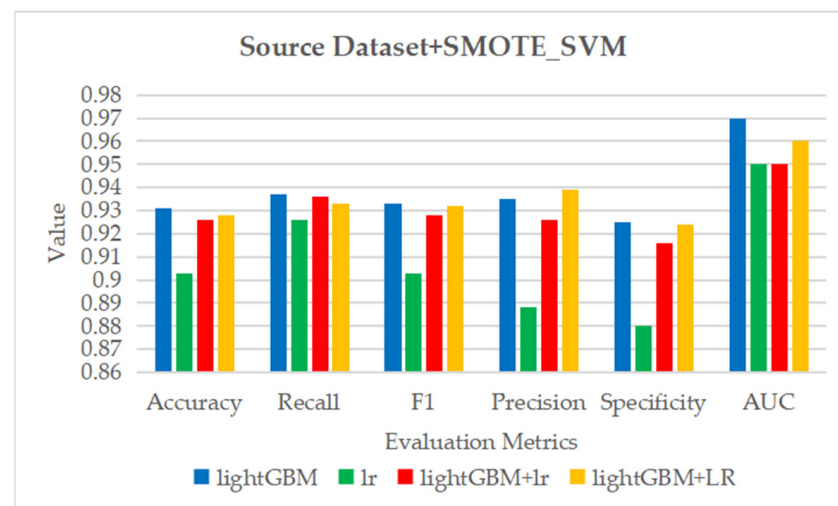
ability of lightGBM and lr model without feature combination processing separately, the classification ability of lightGBM model is stronger. Figure 11 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset balanced by Borderline\_SMOTE. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. The lr model without any feature processing in the four models obtained the lowest AUC average value on the 10 fold data. After the feature combination processing of lighthGBM algorithm, the AUC value obtained by lr model has been improved.

#### 5.2.4. Results Obtained on Dataset Processed by SMOTE\_SVM

The classification performance of the classifiers on the dataset after SMOTE\_SVM processing is described in this part. The balanced dataset processed by SMOTE\_SVM contains the same number of CAD class instances and normal class instances. Each class includes 216 samples, and the entire dataset contains 432 sample instances. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classifiers on the dataset processed by SMOTE\_SVM are displayed in Table 9. Figure 12 is the visual presentation of Table 9. Figure 13 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by SMOTE\_SVM and the AUC value corresponding to each fold ROC curve.

**Table 9.** The average results of 10 tests obtained by classification models on the dataset processed by SMOTE\_SVM.

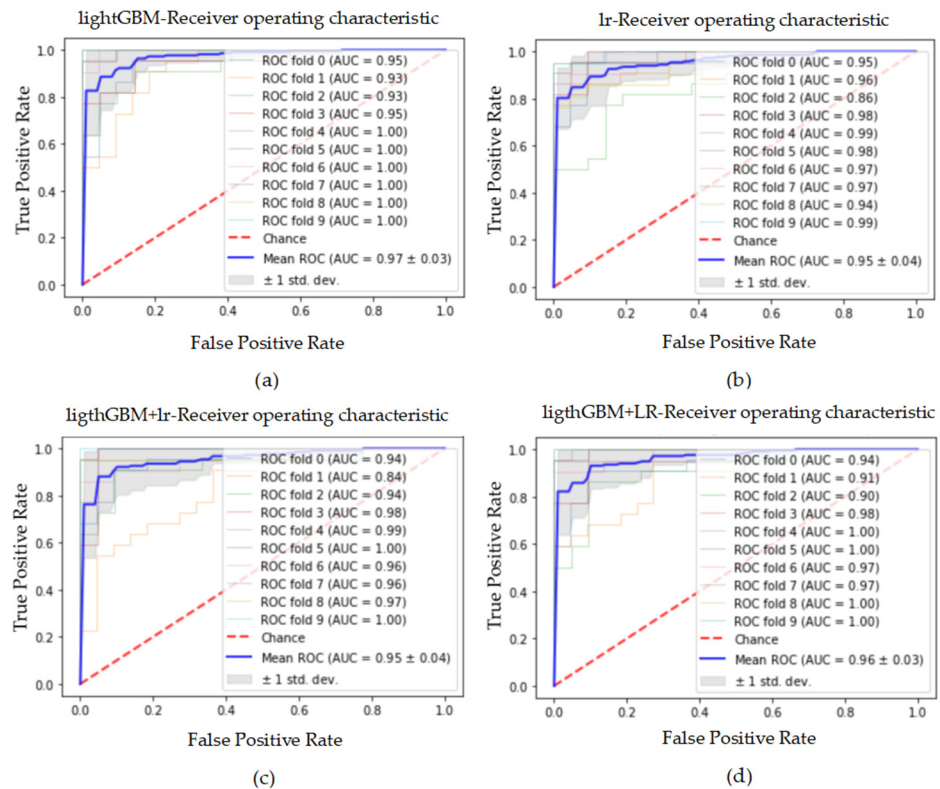
Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.931 ± 0.052	0.937 ± 0.084	0.933 ± 0.046	0.935 ± 0.043	0.925 ± 0.044	0.97 ± 0.03
lr	0.903 ± 0.059	0.926 ± 0.087	0.903 ± 0.055	0.888 ± 0.068	0.880 ± 0.062	0.95 ± 0.04
lightGBM + lr	0.926 ± 0.059	0.936 ± 0.087	0.928 ± 0.052	0.926 ± 0.043	0.916 ± 0.050	0.95 ± 0.04
lightGBM + LR	0.928 ± 0.062	0.933 ± 0.098	0.932 ± 0.054	0.939 ± 0.047	0.924 ± 0.049	0.96 ± 0.03



**Figure 12.** The histogram of Table 9.

It can be seen from Table 9 and Figure 12 that when compared with Table 6 and Figure 6, the classification performance of the models on the dataset processed by the SMOT\_SVM method is significantly improved, mainly reflected in accuracy, recall, specificity and AUC evaluation indicators. On the dataset used in this section, the feature combination method proposed by us has significantly improved the classification ability of lr model. However, the improved results do not exceed the results obtained by the lightGBM model. Therefore, on the dataset studied in this part, lightGBM classifier produced the best classification

results, with accuracy of 93.1%, recall of 93.7%, F1 score of 93.3%, precision of 93.5%, specificity of 92.5% and AUC of 0.97. Figure 13 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset balanced by SMOTE\_SVM. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Among the four models, the lr model without any feature processing has the lowest average AUC value on the 10 fold data. On the dataset processed by lighGBM, the AUC value of lr model increases slightly, but it is lower than the AUC value obtained by lighGBM model itself.



**Figure 13.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by SMOTE\_SVM. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

5.2.5. Results Obtained on Dataset Processed by SMOTE\_Tomek

The dataset used in this section is processed by SMOTE\_Tomek method. The dataset consists of 390 sample instances, among them 195 cases belong to CAD class and 195 cases belong to normal class. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on the dataset processed by SMOTE\_Tomek are expressed in Table 10. Figure 14 is an intuitive display of Table 10. Figure 15 shows the ROC curve of the classification model on the 10-fold test set of the dataset processed by SMOTE\_Tomek and the AUC value corresponding to each fold ROC curve.

**Table 10.** The average results of 10 tests obtained by classification models on the dataset processed by SMOTE\_Tomek.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.946 ± 0.044	0.950 ± 0.065	0.948 ± 0.040	0.948 ± 0.040	0.942 ± 0.044	0.97 ± 0.03
lr	0.926 ± 0.051	0.943 ± 0.083	0.927 ± 0.044	0.917 ± 0.043	0.904 ± 0.046	0.96 ± 0.04
lightGBM + lr	0.946 ± 0.049	0.951 ± 0.071	0.948 ± 0.045	0.949 ± 0.040	0.941 ± 0.045	0.97 ± 0.04
lightGBM + LR	0.941 ± 0.038	0.954 ± 0.064	0.942 ± 0.034	0.933 ± 0.034	0.925 ± 0.036	0.97 ± 0.04

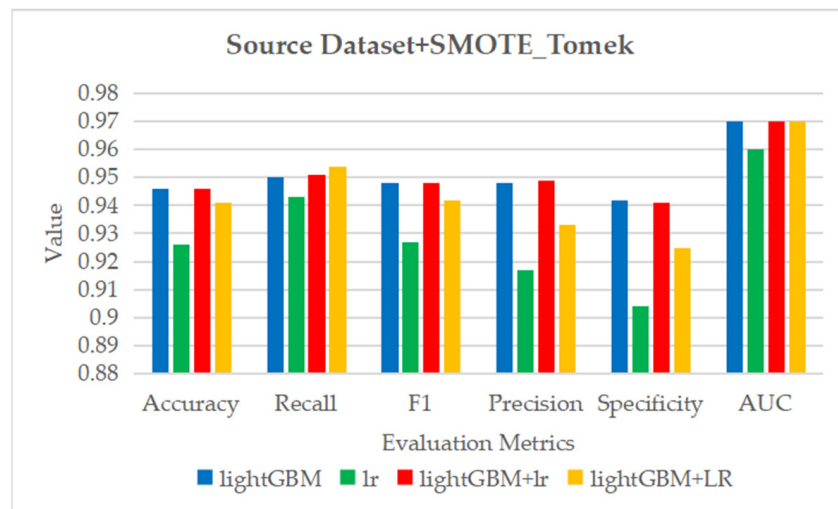


Figure 14. The histogram of Table 10.

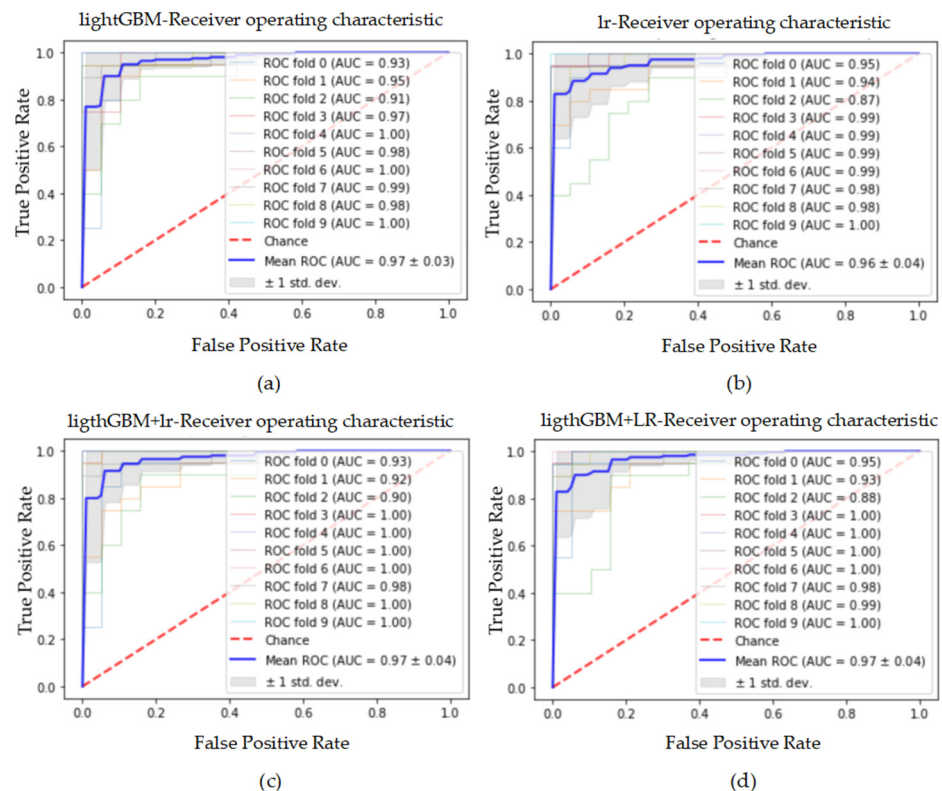


Figure 15. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by SMOTE\_Tomek. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

The results of Tables 6 and 10 show that the classification performance of the four models on the balanced dataset processed by SMOTE\_Tomek method has been greatly improved in addition to the precision index. The best classification results are generated by lr model on the feature set combined by lightGBM, and the best accuracy, recall, F1 score, precision, specificity and AUC are 94.6%, 95.1%, 94.8%, 94.9%, 94.1% and 0.97, respectively. In particular, the highest recall in this section appears in the lightGBM + LR model, which is 95.4%. Similarly, our proposed feature combination method has significantly improved the classification ability of lr model. Figure 15 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model

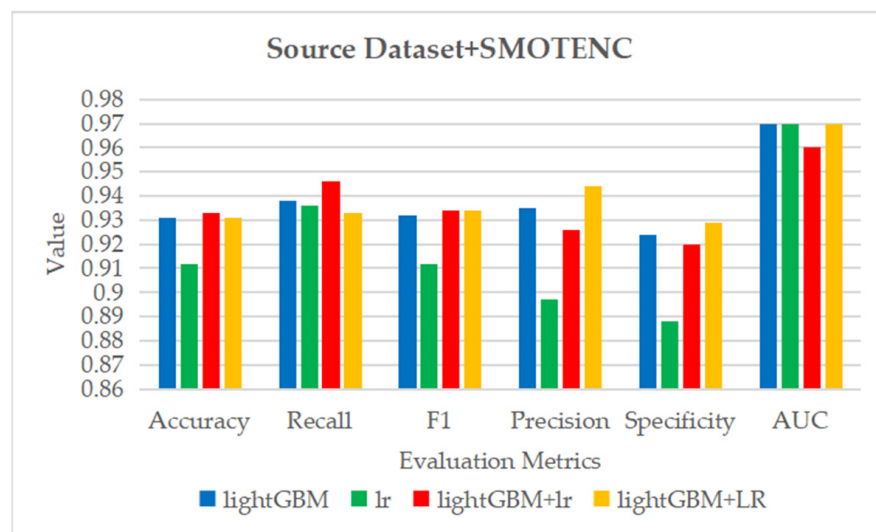
on the dataset balanced by SMOTE\_Tomek. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Moreover, the average AUC value obtained by lr model on the dataset without feature combination processing is the lowest. However, on the dataset processed by lighthGBM model, the AUC result of lr model is improved.

5.2.6. Results Obtained on Dataset Processed by SMOTENC

The dataset processed by the SMOTENC method in this section contains 432 sample instances, of which 216 are classified as CAD class and 216 as normal class. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC of the classification models on the dataset processed by SMOTENC are shown in Table 11. Figure 16 is a visual representation of Table 11. Figure 17 shows the ROC curve of the classification model on the 10-fold test set of the dataset processed by SMOTENC and the AUC value corresponding to each fold ROC curve.

**Table 11.** The average results of 10 tests obtained by classification models on the dataset processed by SMOTENC.

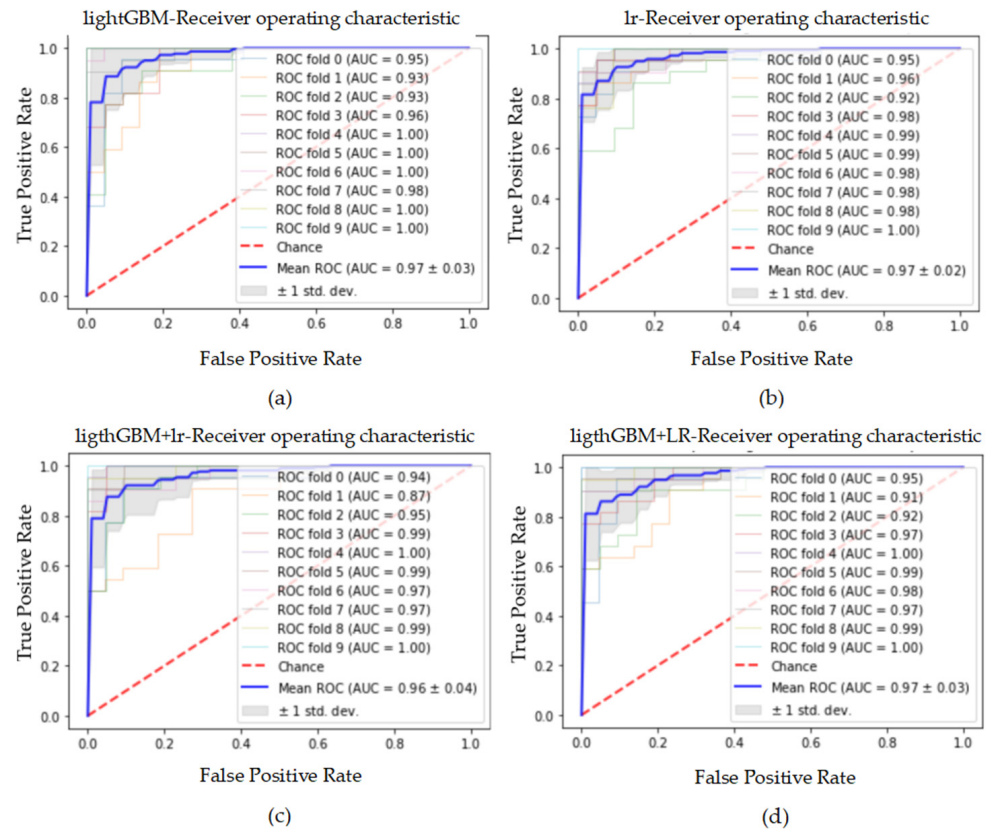
Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.931 ± 0.040	0.938 ± 0.080	0.932 ± 0.035	0.935 ± 0.052	0.924 ± 0.043	0.97 ± 0.03
lr	0.912 ± 0.051	0.936 ± 0.087	0.912 ± 0.046	0.897 ± 0.060	0.888 ± 0.054	0.97 ± 0.02
lightGBM + lr	0.933 ± 0.048	0.946 ± 0.075	0.934 ± 0.044	0.926 ± 0.037	0.920 ± 0.039	0.96 ± 0.04
lightGBM + LR	0.931 ± 0.052	0.933 ± 0.092	0.934 ± 0.044	0.944 ± 0.046	0.929 ± 0.038	0.97 ± 0.03



**Figure 16.** The histogram of Table 11.

Comparing Tables 6 and 11, it can be seen that on the dataset processed by the SMOTENC method, the performance of the classification models in accuracy, recall, specificity and AUC has been improved. The feature combination function of lightGBM model can significantly improve the classification performance of lr model and produce the best classification results. Overall, the best classification results are generated by lightGBM + LR model. The best accuracy, recall, F1 score, precision, specificity and AUC are 93.1%, 93.3%, 93.4%, 94.4%, 92.9% and 0.97, respectively. However, it is worth noting that the highest accuracy and recall of this section appear in the lightGBM + lr model, with the highest accuracy of 93.3% and the highest recall of 94.6%. Figure 17 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset balanced by SMOTENC. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on

each fold test set are different. In addition, the lr model obtained the highest average AUC value on the dataset without feature combination processing. Additionally, the AUC result obtained by lr model on the dataset processed by lighthGBM is reduced.



**Figure 17.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by SMOTENC. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

### 5.2.7. Results Obtained on Standardized Dataset

This section reports the performance results of the classification models used for CAD prediction in terms of accuracy, recall, precision, F1 score, specificity and AUC on the standardized dataset. Data standardization only eliminates the dimension of continuous features, and does not change the size of the dataset. Therefore, there is still a certain class imbalance in the standardized dataset, showing that 71.29% of the samples belong to the CAD class, and 28.71% of the samples belong to the normal class. The average results of 10 tests in terms of accuracy, recall, precision, F1 score, specificity and AUC obtained by the four classification models on the standardized dataset are listed in Table 12. The intuitive exhibition of Table 12 is drawn in Figure 18. Figure 19 shows the ROC curve of the classification model on the 10-fold test set of the standardized dataset and the AUC value corresponding to each fold ROC curve.

**Table 12.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.897 ± 0.049	0.909 ± 0.058	0.931 ± 0.030	0.958 ± 0.044	0.869 ± 0.093	0.92 ± 0.06
lr	0.884 ± 0.059	0.900 ± 0.062	0.922 ± 0.037	0.949 ± 0.045	0.844 ± 0.105	0.93 ± 0.06
lightGBM + lr	0.914 ± 0.040	0.922 ± 0.058	0.942 ± 0.025	0.968 ± 0.036	0.895 ± 0.074	0.92 ± 0.06
lightGBM + LR	0.914 ± 0.060	0.931 ± 0.074	0.942 ± 0.038	0.958 ± 0.038	0.871 ± 0.086	0.93 ± 0.06



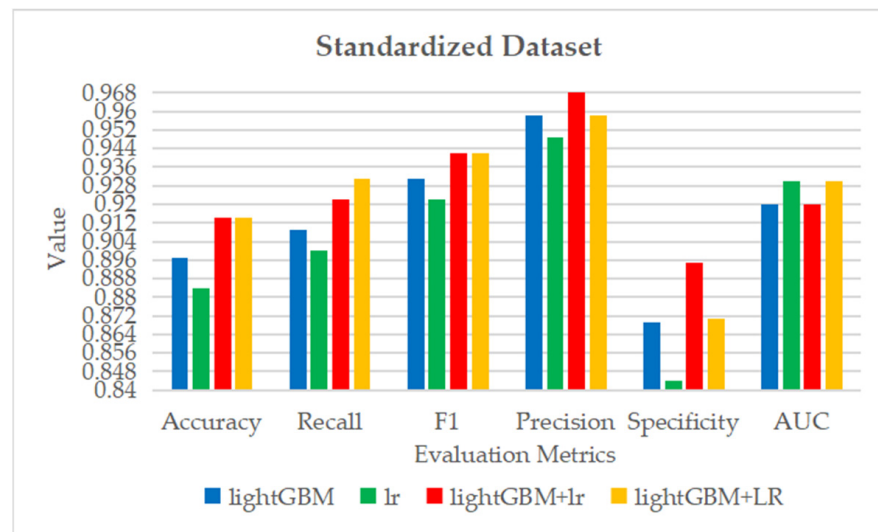


Figure 18. The histogram of Table 12.

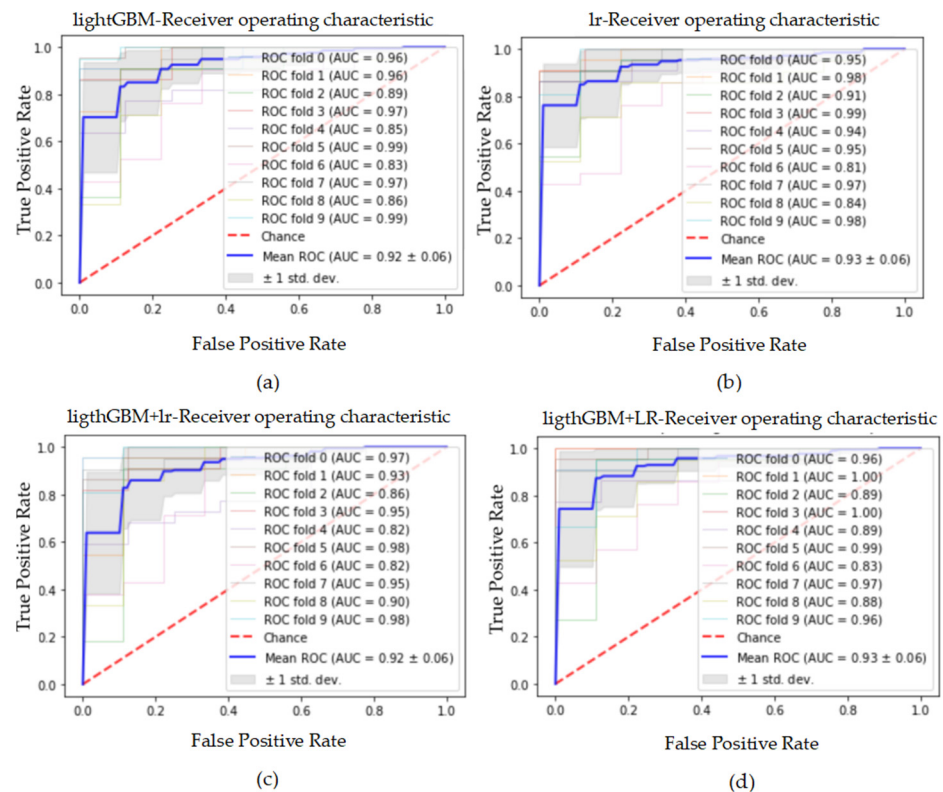


Figure 19. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

By analyzing Tables 6 and 12, we can conclude that on the standardized dataset, the classification performance of lr model has been improved in general. Especially on the dataset processed by lightGBM, lr model obtains the best classification results. The best classification results are accuracy 91.4%, recall 93.1%, F1 score 94.2%, precision 95.8%, specificity 87.1% and AUC 0.93. However, when the lightGBM classifier is applied on a standardized dataset separately, the results are inferior to those obtained on the original dataset. Figure 19 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by

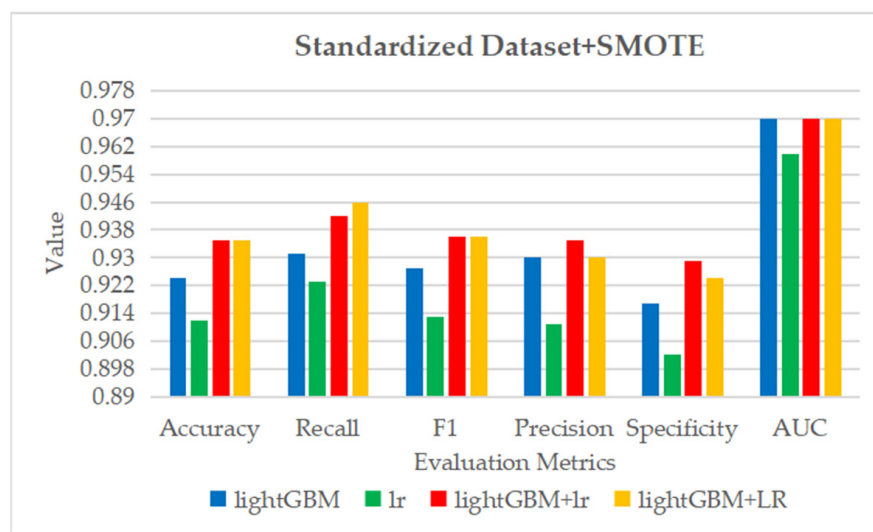
data standardization. It can be seen from the curve that, due to the distribution difference of 10-fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the lr model obtained the highest average AUC value on the dataset without feature combination processing.

#### 5.2.8. Results Obtained on Dataset Processed by Standardization and SMOTE

The research in this section is developed based on the dataset processed by standardization and SMOTE algorithm. The dataset after standardization and SMOTE processing contains 216 samples of CAD class and 216 samples of normal class. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on this dataset are shown in Table 13. Figure 20 is an intuitive display of Table 13. Figure 21 shows the ROC curve of the classification model on the 10-fold test set of the dataset processed by standardization and SMOTE and the AUC value corresponding to each fold ROC curve.

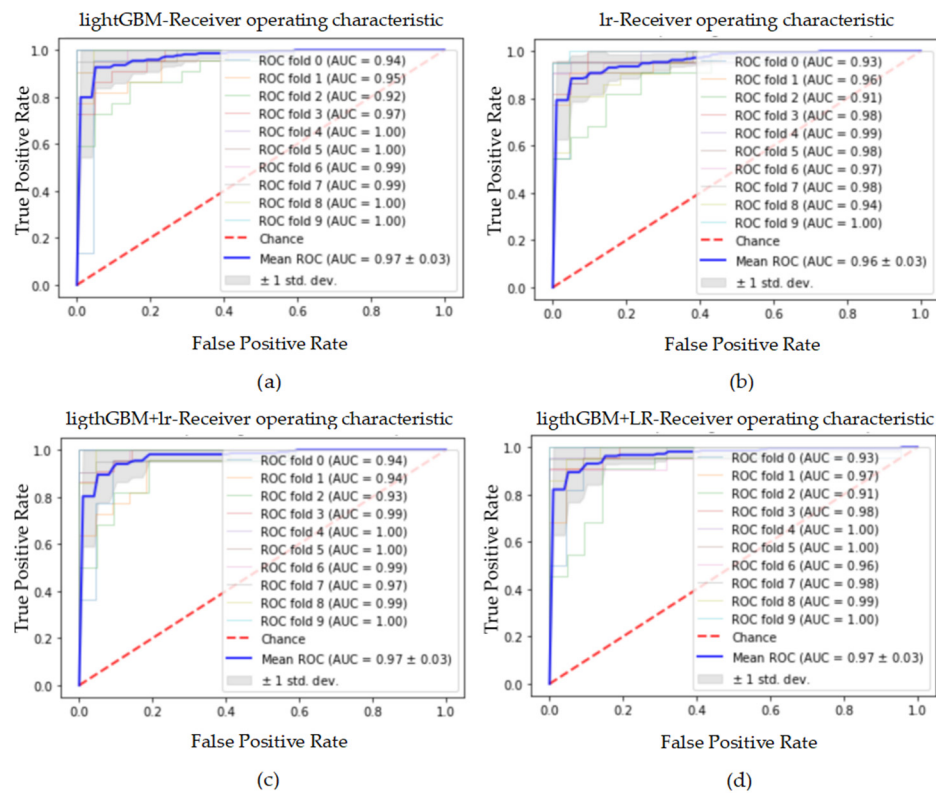
**Table 13.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization and SMOTE.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.924 ± 0.067	0.931 ± 0.100	0.927 ± 0.058	0.930 ± 0.043	0.917 ± 0.048	0.97 ± 0.03
lr	0.912 ± 0.049	0.923 ± 0.081	0.913 ± 0.044	0.911 ± 0.058	0.902 ± 0.053	0.96 ± 0.03
lightGBM + lr	0.935 ± 0.035	0.942 ± 0.066	0.936 ± 0.033	0.935 ± 0.038	0.929 ± 0.035	0.97 ± 0.03
lightGBM + LR	0.935 ± 0.045	0.946 ± 0.071	0.936 ± 0.042	0.930 ± 0.049	0.924 ± 0.049	0.97 ± 0.03



**Figure 20.** The histogram of Table 13.

It can be seen from Table 13 and Figure 20 that when compared with Table 12 and Figure 18, the performance of the classification models on the dataset processed by SMOTE has been significantly improved, mainly reflected in accuracy, recall, specificity and AUC indicators. The best classification performance is obtained by lr model on the dataset combined by lightGBM, with accuracy of 93.5%, recall of 94.6%, F1 score of 93.6%, precision of 93.0%, specificity of 92.4% and AUC of 0.97. Figure 21 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by data standardization and SMOTE. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the proposed feature combination method has improved the AUC value of lr model on the dataset.



**Figure 21.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization and SMOTE. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

5.2.9. Results Obtained on Dataset Processed by Standardization and Borderline\_SMOTE

In this section, the Borderline\_SMOTE class balancing method is applied to the standardized dataset to obtain a new dataset. Four classifiers, namely lightGBM, lr, lightGBM + lr and lightGBM + LR, are applied on this dataset. The dataset processed by standardization and Borderline\_SMOTE contains 432 samples, with CAD class and normal class accounting for 50 %, respectively. The average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on dataset used in this part are displayed in Table 14. Figure 22 is a visual representation of Table 14. Figure 23 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by standardization and Borderline\_SMOTE and the AUC value corresponding to each fold ROC curve.

**Table 14.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization and Borderline\_SMOTE.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.935 ± 0.048	0.944 ± 0.079	0.937 ± 0.044	0.935 ± 0.048	0.926 ± 0.044	0.97 ± 0.03
lr	0.914 ± 0.052	0.934 ± 0.081	0.914 ± 0.048	0.902 ± 0.065	0.895 ± 0.064	0.96 ± 0.02
lightGBM + lr	0.940 ± 0.036	0.957 ± 0.054	0.939 ± 0.035	0.926 ± 0.048	0.923 ± 0.047	0.97 ± 0.03
lightGBM + LR	0.947 ± 0.036	0.961 ± 0.054	0.947 ± 0.034	0.935 ± 0.043	0.932 ± 0.042	0.97 ± 0.03

The comparison of Tables 12 and 14 shows that Borderline\_SMOTE significantly improves the ability of the classifiers. Specifically, in addition to precision, the performance of the four classifiers in other evaluation indicators is significantly improved. The best results consistently appear in the lightGBM + LR classification model. The best accuracy, recall, F1 score, precision, specificity and AUC are 94.7%, 96.1%, 94.7%, 93.5%, 93.2% and 0.97, respectively. From the results, it can be concluded that our proposed feature combination method

has significantly improved the classification performance of lr model. Figure 23 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by data standardization and Borderline\_SMOTE. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the proposed feature combination method has improved the AUC value of lr model on this dataset.

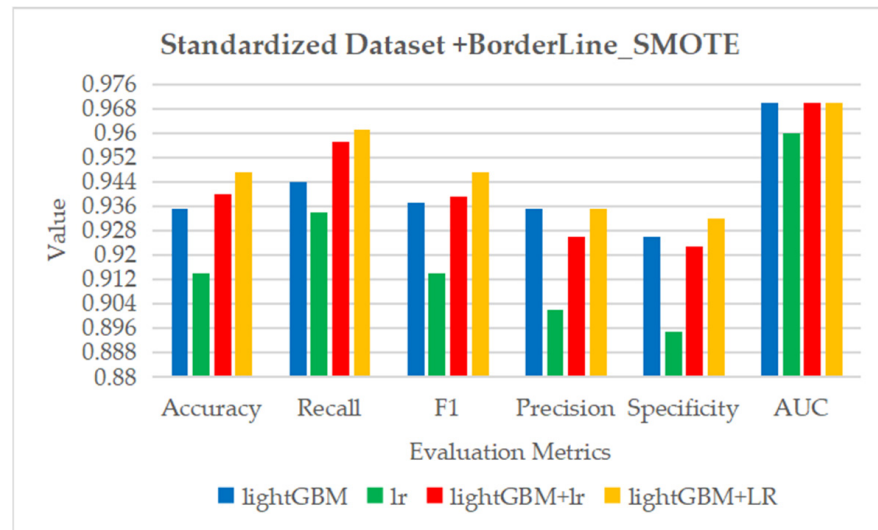


Figure 22. The histogram of Table 14.

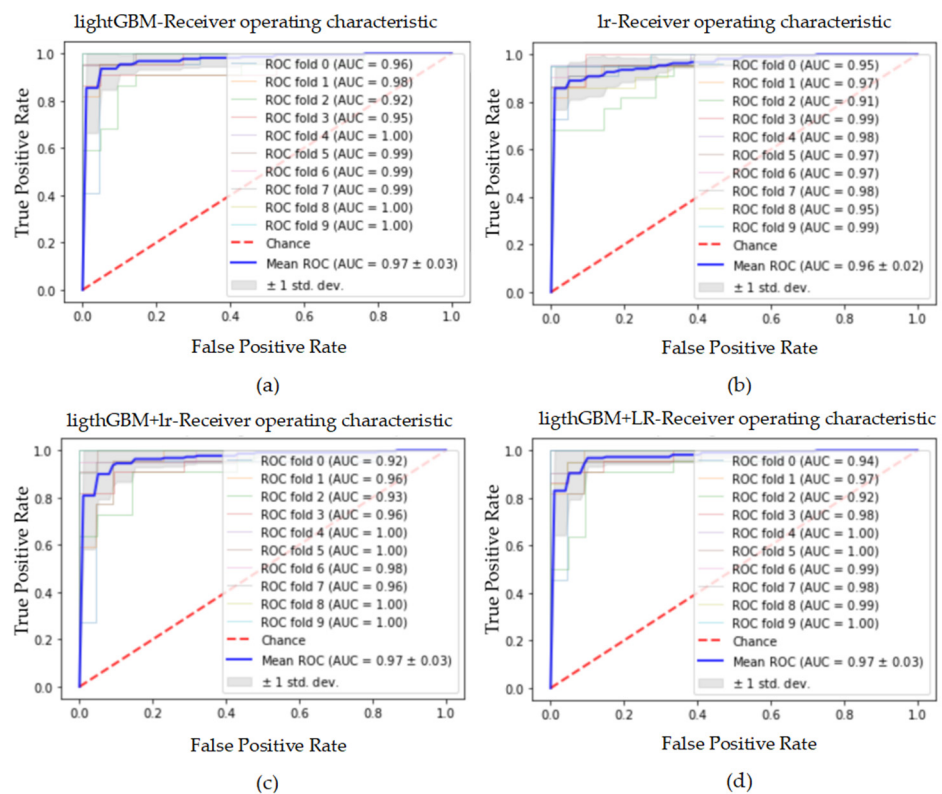


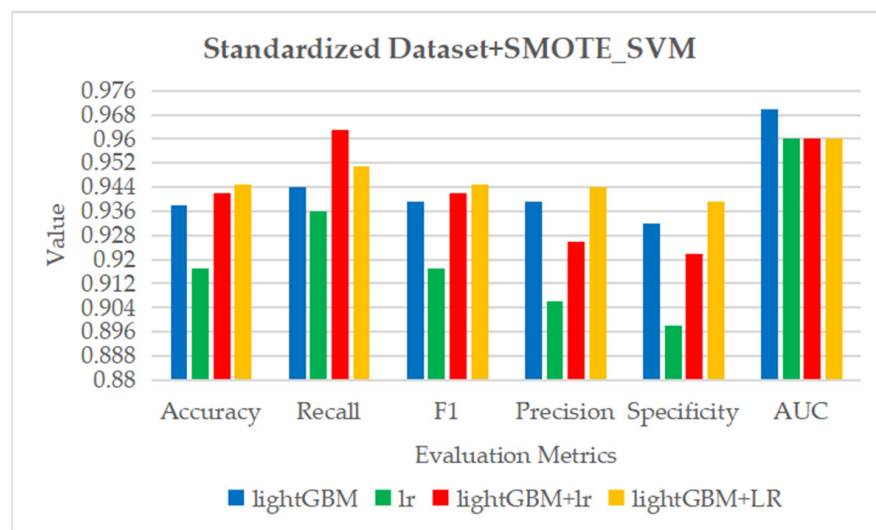
Figure 23. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization and Borderline\_SMOTE. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

5.2.10. Results Obtained on Dataset Processed by Standardization and SMOTE\_SVM

This section shows the performance of classification models combined with data standardization and SMOTE\_SVM processing methods. The number of sample instances in CAD class and normal class in the dataset processed by standardization and SMOTE\_SVM method is the same, and both contain 216 cases. The average results of 10 tests in terms of accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on this dataset are listed in Table 15. Figure 24 is a graphical representation of Table 15. Figure 25 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by standardization and SMOTE\_SVM and the AUC value corresponding to each fold ROC curve.

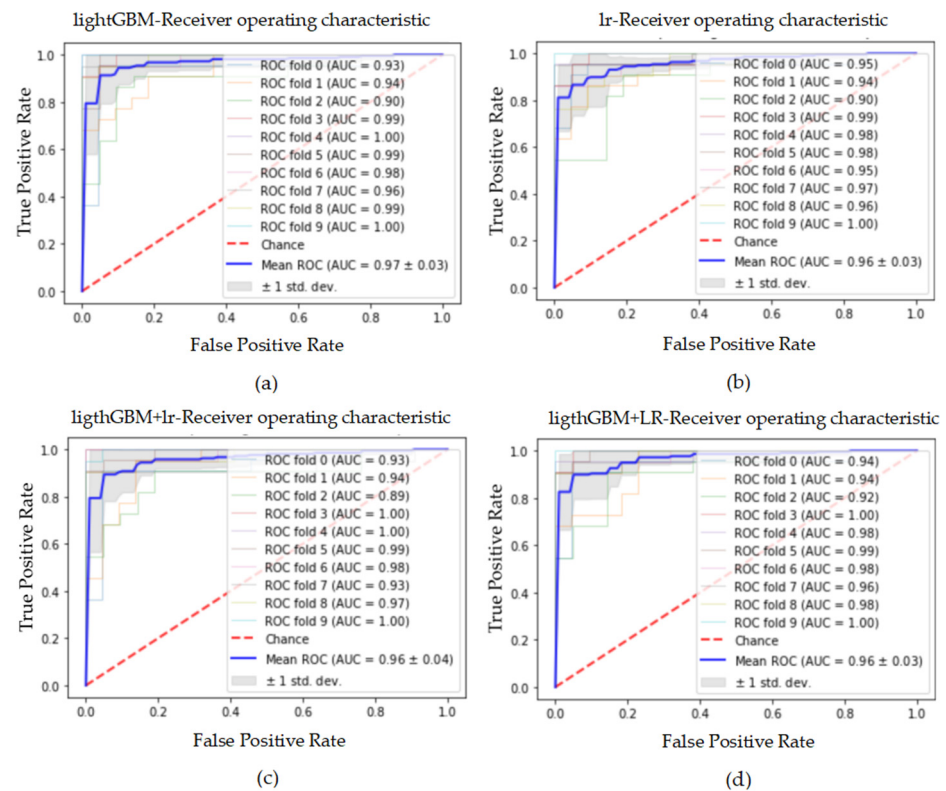
**Table 15.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization and SMOTE\_SVM.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.938 ± 0.058	0.944 ± 0.080	0.939 ± 0.052	0.939 ± 0.052	0.932 ± 0.056	0.97 ± 0.03
lr	0.917 ± 0.047	0.936 ± 0.082	0.917 ± 0.043	0.906 ± 0.061	0.898 ± 0.056	0.96 ± 0.03
lightGBM + lr	0.942 ± 0.040	0.963 ± 0.065	0.942 ± 0.037	0.926 ± 0.038	0.922 ± 0.037	0.96 ± 0.04
lightGBM + LR	0.945 ± 0.036	0.951 ± 0.062	0.945 ± 0.033	0.944 ± 0.041	0.939 ± 0.036	0.96 ± 0.03



**Figure 24.** The histogram of Table 15.

As shown in Table 15 and Figure 24, when compared with Table 12 and Figure 18, it can be found that SMOTE\_SVM processing significantly improves the ability of the classifiers. Specifically, in addition to precision, the performance of the four classifiers in other evaluation indexes is significantly improved. Overall, the best results are produced by the lightGBM + LR model. The best accuracy, recall, F1 score, precision, specificity and AUC are 94.5%, 95.1%, 94.5%, 94.4%, 93.9% and 0.96, respectively. However, on the dataset used in this section, the highest recall and AUC appear in lightGBM + lr model and the lightGBM model, respectively. The highest recall and AUC are 96.3% and 0.97. It can be seen that the feature combination of lightGBM model can significantly improve the classification ability of lr model. Figure 25 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by data standardization and SMOTE\_SVM. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the proposed feature combination method does not improve the AUC value of lr model on this dataset.



**Figure 25.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization and SMOTE\_SVM. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

5.2.11. Results Obtained on Dataset Processed by Standardization and SMOTE\_Tomek

In this section, we applied SMOTE\_Tomek class balancing method to the standardized dataset, and obtained a balanced dataset. The balanced dataset contains 430 samples, of which 215 belong to CAD class and 215 belong to normal class. Table 16 shows the average results of 10 tests on accuracy, recall, precision, F1 score, specificity and AUC obtained by the classification models on this dataset. Figure 26 is an intuitive display of Table 16. Figure 27 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by standardization and SMOTE\_Tomek and the AUC value corresponding to each fold ROC curve.

**Table 16.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization and SMOTE\_Tomek.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.937 ± 0.048	0.939 ± 0.085	0.940 ± 0.041	0.948 ± 0.039	0.936 ± 0.032	0.98 ± 0.02
lr	0.912 ± 0.051	0.926 ± 0.089	0.913 ± 0.045	0.911 ± 0.066	0.898 ± 0.055	0.96 ± 0.03
lightGBM + lr	0.947 ± 0.050	0.948 ± 0.073	0.948 ± 0.046	0.953 ± 0.047	0.945 ± 0.048	0.98 ± 0.02
lightGBM + LR	0.944 ± 0.048	0.952 ± 0.075	0.946 ± 0.043	0.944 ± 0.046	0.936 ± 0.045	0.97 ± 0.03

By comparing Tables 12 and 16, we can infer that the classification performance of the four models has significantly improved on the dataset balanced by the SMOTE\_Tomek class balancing method. On the dataset used in this section, four models show good classification ability. In particular, on the feature set combined by lightGBM, lr model achieved the best classification results in this study, with accuracy of 94.7%, recall of 94.8%, F1 score of 94.8%, precision of 95.3%, specificity of 94.5% and AUC of 0.98. This shows that the method of using the feature combination function of lightGBM model to improve the classification

ability of lr model is effective. This also fully proves the effectiveness of our proposed method. Figure 27 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by data standardization and SMOTE\_Tomek. It can be seen from the curve that, due to the distribution difference of 10 fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the proposed feature combination method has significantly improved the AUC value of lr model on this dataset. Additionally, the highest AUC value of this study was produced.

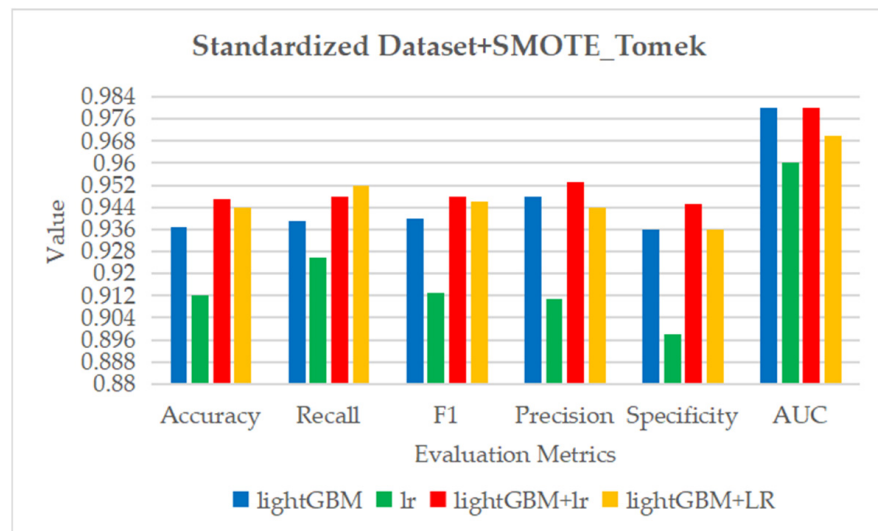


Figure 26. The histogram of Table 16.

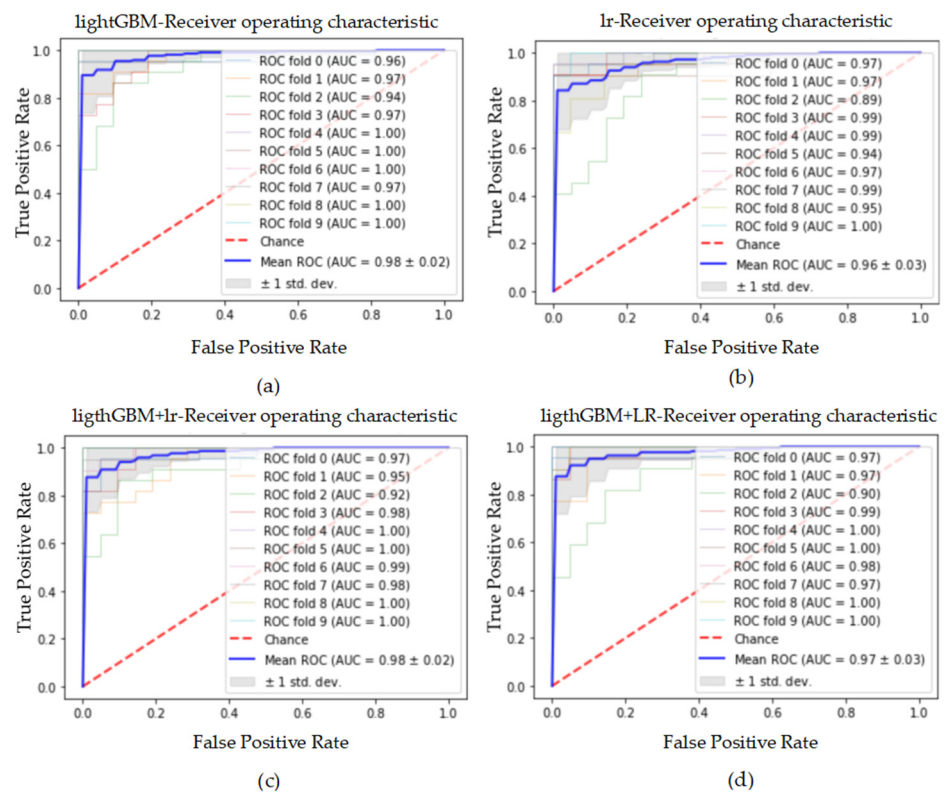


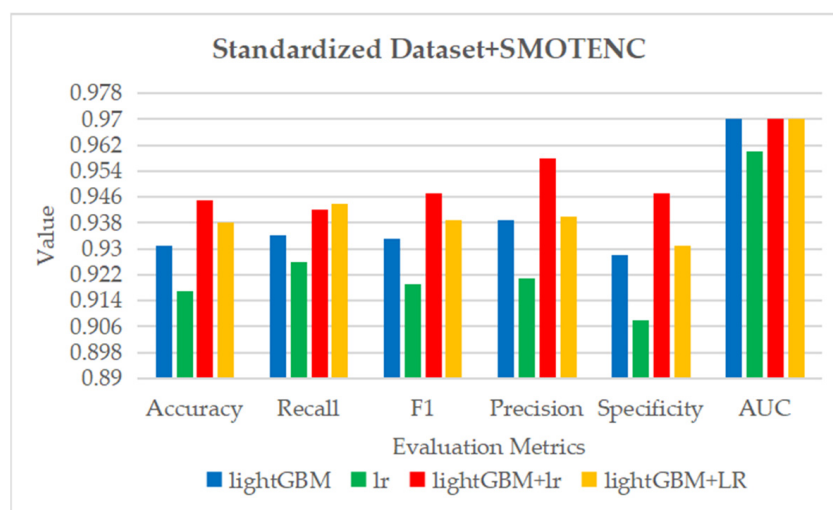
Figure 27. The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization and SMOTE\_Tomek. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

### 5.2.12. Results Obtained on Dataset Processed by Standardization and SMOTENC

The performance results of the four classification models for CAD prediction in terms of accuracy, recall, precision, F1 score, specificity and AUC on the dataset processed by data standardization and SMOTENC method are reported in this section. The dataset used in this part consists of 432 samples, including 216 instances of CAD class and 216 instances of normal class. Table 17 exhibits the average results of 10 tests obtained by the four classification models on this dataset. Figure 28 is a histogram of Table 17. Figure 29 shows the ROC curve of the classification model on the 10 fold test set of the dataset processed by standardization and SMOTENC and the AUC value corresponding to each fold ROC curve.

**Table 17.** The average results of 10 tests obtained by classification models on the dataset processed by data standardization and SMOTENC.

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
lightGBM	0.931 ± 0.058	0.934 ± 0.089	0.933 ± 0.052	0.939 ± 0.056	0.928 ± 0.055	0.97 ± 0.03
lr	0.917 ± 0.046	0.926 ± 0.086	0.919 ± 0.042	0.921 ± 0.061	0.908 ± 0.052	0.96 ± 0.03
lightGBM + lr	0.945 ± 0.050	0.942 ± 0.080	0.947 ± 0.045	0.958 ± 0.049	0.947 ± 0.047	0.97 ± 0.03
lightGBM + LR	0.938 ± 0.043	0.944 ± 0.074	0.939 ± 0.040	0.940 ± 0.055	0.931 ± 0.050	0.97 ± 0.02



**Figure 28.** The histogram of Table 17.

It can be seen from Table 17 and Figure 28, compared with Table 12 and Figure 18, SMOTENC method improves the performance of the classification models in terms of accuracy, recall, specificity and AUC. On the dataset used in this section, the feature combination of lightGBM also has significantly improved the classification ability of lr model. The lightGBM + lr model obtained the best results with accuracy of 94.5%, recall of 94.2%, F1 score 94.7%, precision of 95.8%, specificity of 94.7% and AUC of 0.97. Figure 29 shows the ROC curve and corresponding AUC value of each fold in the 10-fold cross-validation obtained by classification model on the dataset processed by data standardization and SMOTENC. It can be seen from the curve that, due to the distribution difference of 10-fold data, the AUC values obtained by the model on each fold test set are different. Additionally, the proposed feature combination method has significantly improved the AUC value of lr model on this dataset.

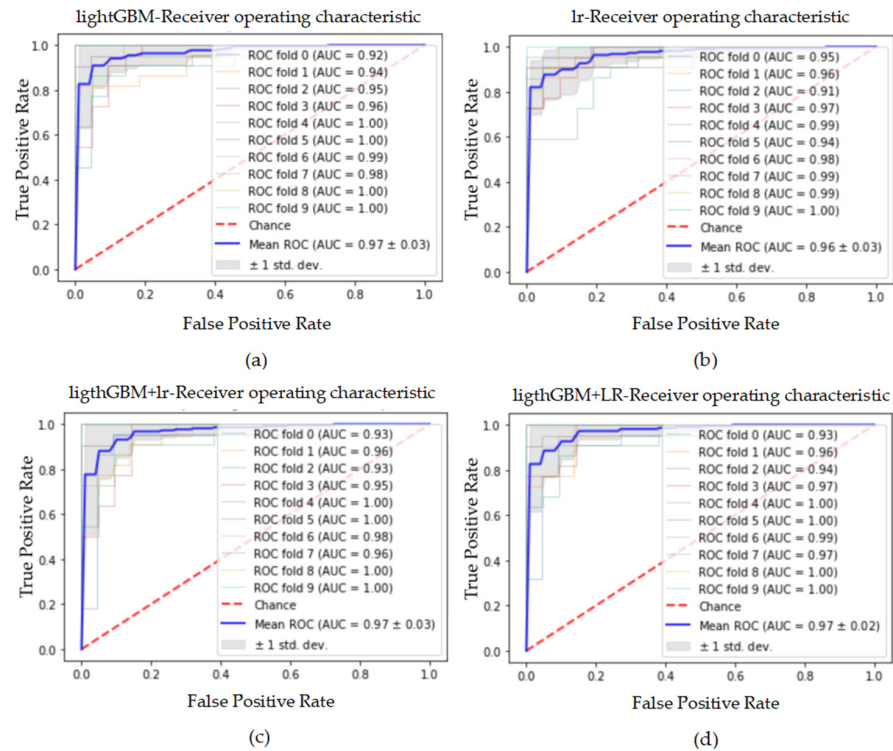
### 5.3. Results Analysis

#### 5.3.1. Analysis of the Influence of Class Balancing Methods on the Performance of Classification Models

This section reports the impact trend of five class balancing methods on the performance of classification models. In order to eliminate the impact of the data standardization process-



ing method on the performance of the model, we analyze it based on the same basic dataset and the same feature processing. Therefore, study of this section is divided into two parts, one is based on the original dataset, and the other is based on the standardized dataset.



**Figure 29.** The ROC curve and AUC value of each fold in the 10-fold cross-validation obtained by classification models on the dataset processed by data standardization and SMOTENC. (a–d) are the ROC curves of lightGBM, lr, lightGBM + lr and lightGBM + LR classifiers, respectively.

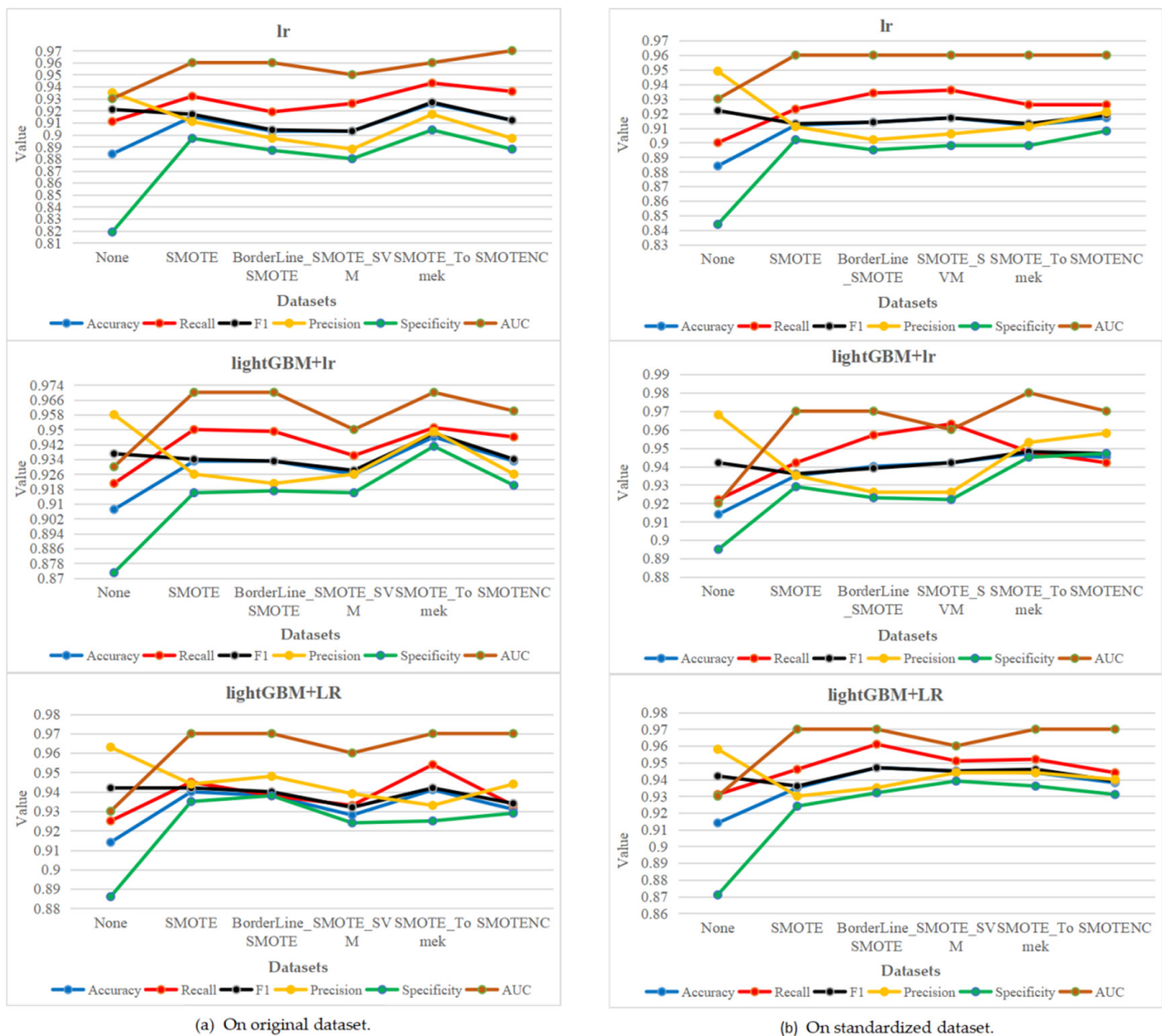
### 1. Research Based on Original Dataset

In this part, we describe in detail the change trend of the performance of the classification models after applying five class balancing methods on the original dataset, as shown in Figure 30a.

The three child figures from top to bottom in Figure 30a are the trend charts of the performance evaluation indexes of the three classification models with the five class balancing methods, respectively. The three classification models are lr, lightGBM + lr and lightGBM + LR. The mark ‘None’ in Figure 30 refers to the dataset that does not apply the class balancing methods. The marks ‘SMOTE’, ‘BorderLine\_SMOTE’, ‘SMOTE\_SVM’, ‘SMOTE\_Tomek’ and ‘SMOTENC’ correspond to the datasets that applied the corresponding class balancing methods, respectively.

By analyzing the results of three classification models on the ‘None’ dataset, it can be found that the distribution of results is scattered, that is, the model obtains good results on some indicators, but it is poor on other indicators. This may be related to the skewness and class imbalance of the original dataset. After applying the class balancing methods, the results of the model are improved, and the distribution of the results tends to be concentrated. For lr and lightGBM + lr classification models, the most concentrated and best results appear on the dataset balanced by SMOTE\_Tomek method. For the lightGBM + LR model, the most concentrated results appear on the dataset balanced by the BorderLine\_SMOTE method, and the best results also appear on the dataset balanced by the SMOTE\_Tomek method. The reason for this phenomenon in the lightGBM + LR model may be related to the addition of the original dataset to the recombined dataset. The change trend of the results of the three classification models in Figure 30a confirms that the five class balancing methods can effectively improve the classification performance and

stability of the model on the original dataset. Especially, SMOTE\_Tomek algorithm is a more effective method.



**Figure 30.** The trend charts of the performance evaluation indexes with class balancing methods on original dataset and standardized dataset.

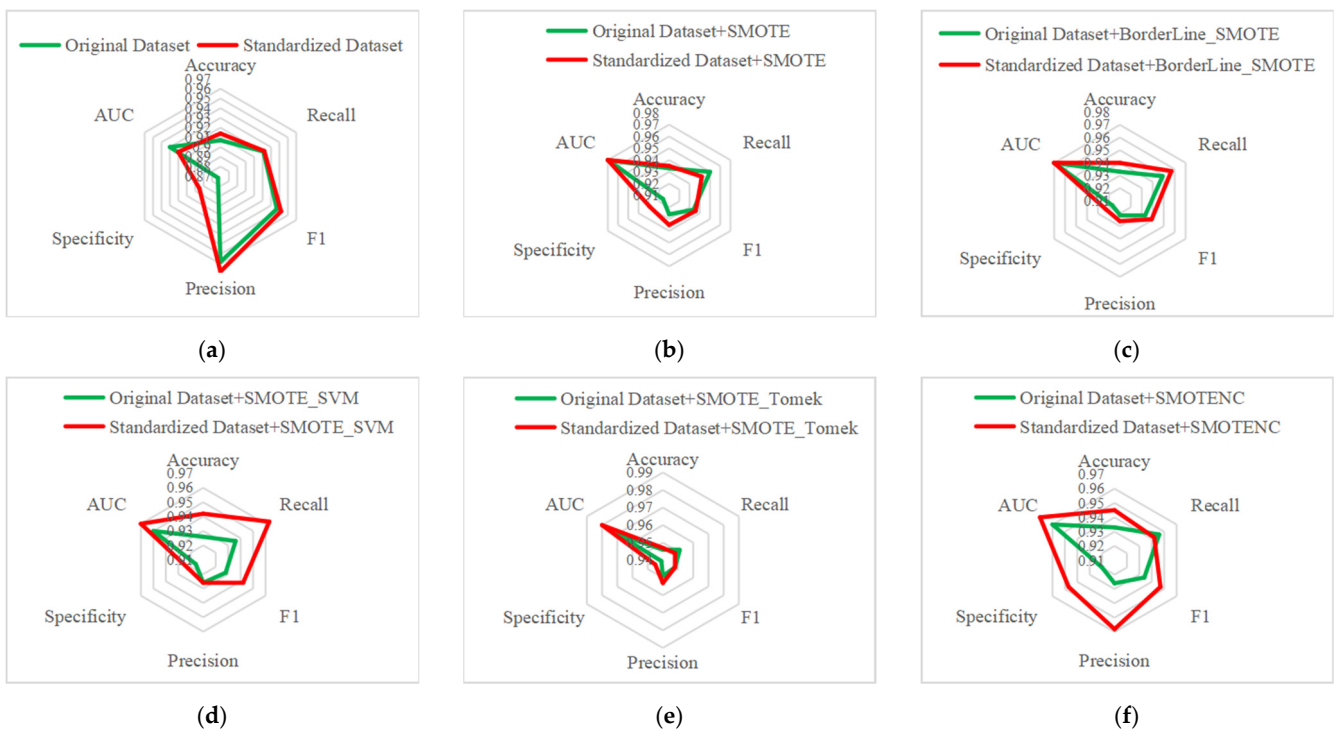
## 2. Research Based on Standardized Dataset

The change trend of the performance of the three classification models after applying five class balancing methods on the standardized dataset are described in Figure 30b.

By analyzing Figure 30b, we can find that after applying the class balancing methods, the results of the models are improved, and the distribution of the results tends to be concentrated. For lightGBM + lr model, the most concentrated and best results appear on the dataset balanced by SMOTE\_Tomek method. For the lr model, the most concentrated results appear on the dataset balanced by the SMOTE\_Tomek method, and the best results appear on the dataset balanced by the SMOTE\_SVM method. For the lightGBM + LR model, the most concentrated results appear on the dataset balanced by the SMOTE\_SVM method, and the best results appear on the dataset balanced by the BorderLine\_SMOTE method. The variation trend of the results of the three classification models in Figure 30b confirms that on the standardized dataset, the five class balancing methods can effectively improve the classification performance and stability of the model.

### 5.3.2. Analysis of the Influence of Data Standardization Method on the Performance of Classification Models

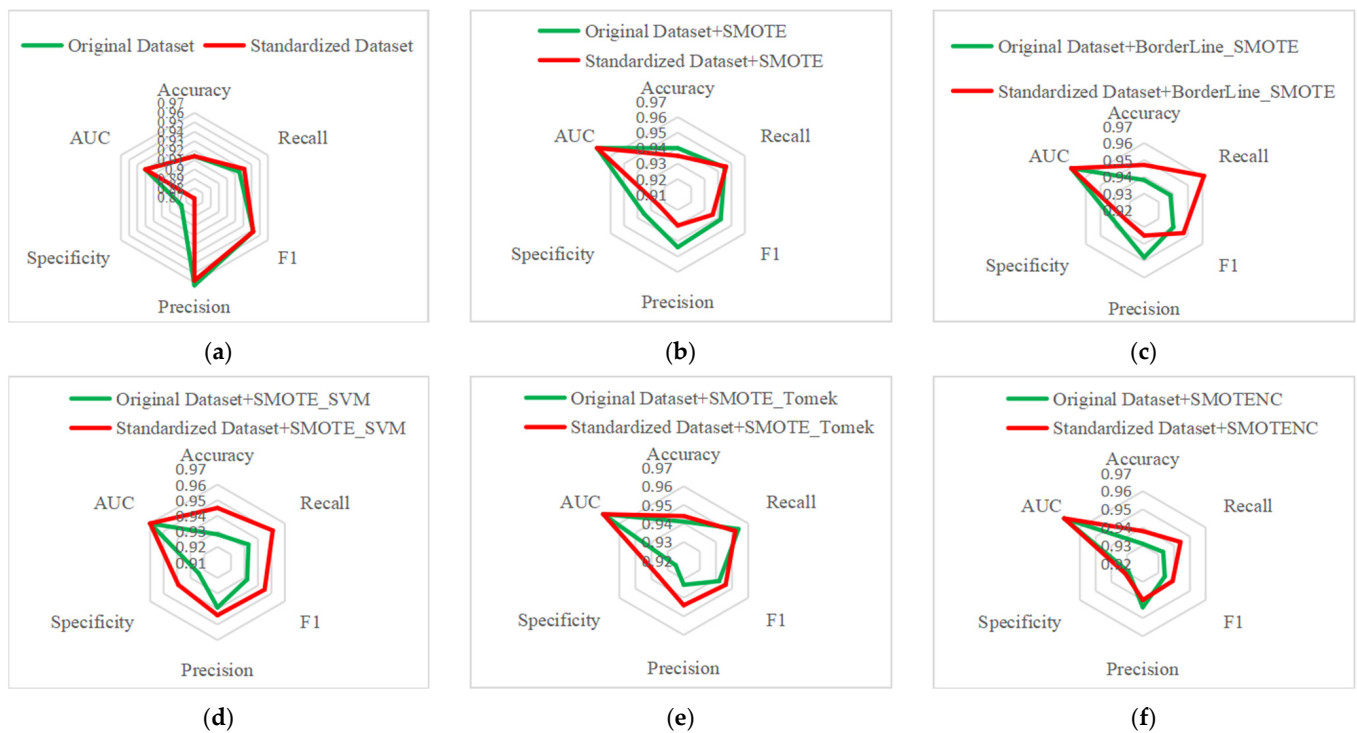
In this part, we analyze the impact trend of data standardization method on the performance of classification models. According to the foregoing, the best classification results on each dataset often come from the lightGBM + lr model or the lightGBM + LR model. Therefore, this part mainly analyzes the influence of data standardization method on lightGBM + lr model and lightGBM + LR model. As shown in Figures 31 and 32, respectively. The study of this part adopts the matching method, that is, the application of other processing methods is the same for the two sets of datasets used for comparison, except whether data standardization processing method is applied or not. Therefore, each model has six groups of comparative datasets, which are the datasets without balancing and the datasets processed by SMOTE, BorderLine\_SMOTE, SMOTE\_SVM, SMOTE\_Tomek and SMOTENC, respectively. Figure 31 shows the performance comparison of lightGBM + lr model on six groups of datasets. Figure 32 shows the performance comparison of lightGBM + LR model on six groups of datasets. The (a) to (f) of Figures 31 and 32 correspond to the datasets without balancing and the datasets processed by SMOTE, BorderLine\_SMOTE, SMOTE\_SVM, SMOTE\_Tomek and SMOTENC, respectively.



**Figure 31.** The performance comparison of lightGBM + lr model on six groups of datasets. (a–f) correspond to the datasets without balancing and the datasets processed by SMOTE, BorderLine\_SMOTE, SMOTE\_SVM, SMOTE\_Tomek and SMOTENC, respectively.

From the 6 groups comparison diagram of (a)–(f) in Figure 31, it can be found that the standardization of datasets can significantly improve the classification ability of lightGBM + lr model.

From the 6 groups comparison diagram of (a)–(f) in Figure 32, it can be found that, on the whole, the data standardization method can significantly improve the classification ability of lightGBM + LR model. Although the data standardization method does not improve the performance of lightGBM + LR model on the dataset processed by SMOTE method, the recognition ability of lightGBM + LR model to positive samples, namely the recall index, has been improved on the standardized dataset, which is also considered to be very important.



**Figure 32.** The performance comparison of lightGBM + LR model on six groups of datasets. (a–f) correspond to the datasets without balancing and the datasets processed by SMOTE, BorderLine\_SMOTE, SMOTE\_SVM, SMOTE\_Tomek and SMOTENC, respectively.

### 5.3.3. Analysis of Ablation Study

The composition of the CAD prediction model studied in this paper can be divided into four modules, namely data preprocessing, class balance processing, feature combination and class prediction. The modules of data preprocessing, class balance processing and feature combination are designed to optimize the performance of classification algorithm. To understand which modules are critical to detection performance, we analyzed the ablation study of the model. That is, after removing one module in turn, the change of the model’s detection performance will be observed. This can intuitively reflect the impact of the module on the classification performance of model. In this study, five different class balancing algorithms are tried. The experimental results show that the best classification results are obtained by lr model on the dataset processed by SMOTE\_Tomek method and lightGBM algorithm. Therefore, the class balance processing module studied in this part applies SMOTE\_Tomek method, and the feature combination module does not involve the recombination of the original dataset. Tables 18–20 show the classification results of the prediction model after removing the modules of data preprocessing, class balance processing and feature combination in turn. Figure 33 compares the improvement effects of different modules on the performance results of model.

**Table 18.** The classification results of the prediction model after removing the modules of data preprocessing.

Modules	Accuracy	Recall	F1	Precision	Specificity	AUC
ALL	0.947 ± 0.050	0.948 ± 0.073	0.948 ± 0.046	0.953 ± 0.047	0.945 ± 0.048	0.98 ± 0.02
NP	0.946 ± 0.049	0.951 ± 0.071	0.948 ± 0.045	0.949 ± 0.040	0.941 ± 0.045	0.97 ± 0.04
P	0.001	−0.003	0.000	0.004	0.004	0.01

NP, no data preprocessing. P, data preprocessing.

**Table 19.** The classification results of the prediction model after removing the modules of class balance processing.

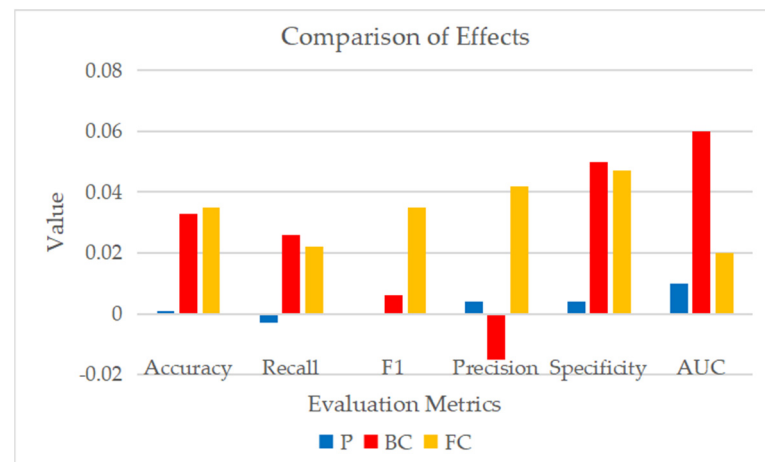
Modules	Accuracy	Recall	F1	Precision	Specificity	AUC
ALL	0.947 ± 0.050	0.948 ± 0.073	0.948 ± 0.046	0.953 ± 0.047	0.945 ± 0.048	0.98 ± 0.02
NBC	0.914 ± 0.040	0.922 ± 0.058	0.942 ± 0.025	0.968 ± 0.036	0.895 ± 0.074	0.92 ± 0.06
BC	0.033	0.026	0.006	−0.015	0.050	0.06

NBC, no methods of balancing classes. BC, balancing classes.

**Table 20.** The classification results of the prediction model after removing the modules of feature combination.

Modules	Accuracy	Recall	F1	Precision	Specificity	AUC
ALL	0.947 ± 0.050	0.948 ± 0.073	0.948 ± 0.046	0.953 ± 0.047	0.945 ± 0.048	0.98 ± 0.02
NFC	0.912 ± 0.051	0.926 ± 0.089	0.913 ± 0.045	0.911 ± 0.066	0.898 ± 0.055	0.96 ± 0.03
FC	0.035	0.022	0.035	0.042	0.047	0.02

NFC, no feature combination. FC, feature combination.

**Figure 33.** Comparison of improvement effects of different modules on performance results.

By analyzing Tables 18–20, it can be seen that the feature combination module can significantly improve the results of all performance indicators. In addition to the precision, the class balance processing module also shows a significant improvement in other performance indexes. Additionally, the data preprocessing module also shows an improvement in other performance indexes, in addition to the recall and F1 score. Figure 33 intuitively shows the comparison of the three processing modules on the improvement of model prediction performance. It can be seen from Figure 33 that the three processing modules are all effective methods to improve the classification ability of the model, but they have different effects on the performance improvement of the classification model. Among them, the feature combination has the strongest effect on the improvement of model performance. The second is class balance processing. The data preprocessing module has the weakest effect on the improvement of model performance. Specifically, the class balance processing has the strongest ability to improve recall, specificity and AUC. Feature combination processing has the strongest ability to improve accuracy, F1 score and precision. Therefore, the combination of class balance and feature combination is an effective method to realize the early and accurate diagnosis of CAD.

#### 5.3.4. Analysis of Model Loss

In this section, we analyze the loss of the proposed model on the training set and test set. As mentioned earlier, lr model obtains the best classification performance in this study

on the dataset processed by data preprocessing, SMOTE\_Tomek method and lighthGBM algorithm. This model is the complete embodiment of our proposed method. This section is analyzed and discussed based on this model, namely lighthGBM + lr classifier. At the same time, we also calculate the loss of lr model on the dataset without feature combination processing, that is, lr classifier. In the development process of the model, we applied the 10-fold cross-validation technology. Therefore, the loss in each epoch is the average of the 10-fold data. Figures 34 and 35, respectively, show the change trend of the loss of the two models on the training set and the test set with the update of the number of epochs. It can be seen from Figures 34 and 35 that, on the training set, with the increase of epoch times, the loss of both models gradually decreases and finally approaches a stable value. On the test set, with the update of the number of epochs, the loss of lr model gradually decreases and finally tends to be stable. The loss of lightGBM + lr model shows a trend of decreasing first, then increasing finally stabilizing. The lowest loss occurs when epoch = 4. These indicate that the model shows a certain over fitting when epochs increases to 5. Therefore, we take epochs = 4 to predict the test set.

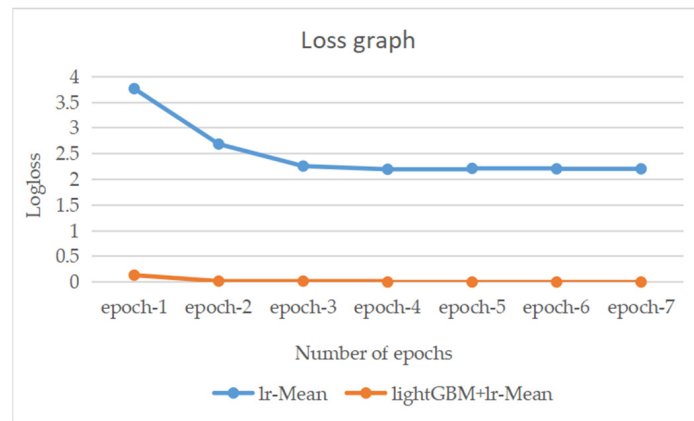


Figure 34. The change trend of the loss of the two models on the training set.

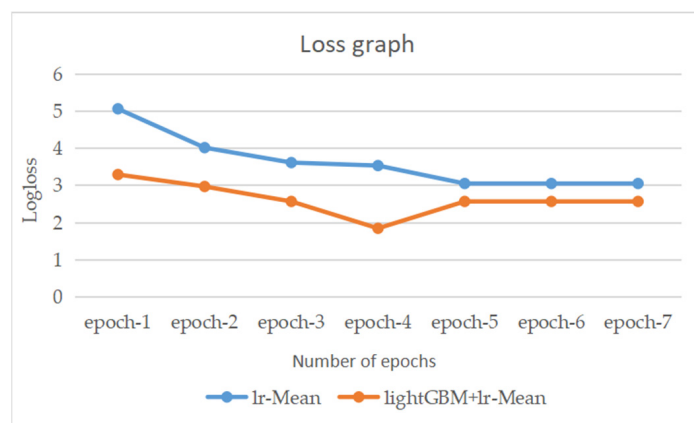


Figure 35. The change trend of the loss of the two models on the test set.

### 6. Discussion

In this study, we proposed a model for early, accurate and rapid detection of CAD. The model is based on the classical logistic regression algorithm. Logistic regression algorithm with good interpretability is the most commonly used research method in medical problems. Especially in medical binary classification problems, logistic regression algorithm has irreplaceable advantages. However, the disadvantage of logistic regression algorithm is also obvious, that is, it needs a lot of manual feature engineering. There is no doubt that this will take a lot of time. Tree model algorithm has been widely used in feature selection

and feature combination, because of its efficient node splitting mechanism. Therefore, in this study, we used the tree model algorithm to realize the feature engineering automation of lr algorithm. At the same time, in order to improve the classification performance of the proposed model, we applied five sampling methods to solve the class imbalance problem in the dataset. The corresponding solutions to the problem of feature distribution skewness in the dataset were proposed. In addition, the 10 fold cross validation technique was used to test the robustness of the model. Seven important evaluation indexes, namely, accuracy, recall, F1 score, precision, specificity, ROC curve and AUC, were used for the evaluation and analysis of the model. The experimental results shown that the data preprocessing, class balance processing and feature combination methods can significantly improve the classification ability of lr model. However, different processing methods have different effects on the classification performance of the model. In general, the feature combination processing based on lighthGBM algorithm has the strongest effect on improving the performance of the model, followed by class balance processing. The data standardization method has the weakest effect on the improvement of model performance. When it comes to the performance evaluation metrics of the model, we can find that class balance processing has the strongest ability to improve recall, specificity and AUC. Feature combination processing based on lighthGBM algorithm has the strongest ability to improve accuracy, F1 score and precision. Besides, the data standardization method does not contribute to the recall rate and F1 score index of the model, but also improved the performance of the model in other indicators. Therefore, the combination of data preprocessing method, class balance method and feature combination method can improve the performance of the classification model in all evaluation metrics. Our experimental results also confirm this. The best classification result of this study is generated by lr model on the dataset after standardization, resampling and feature combination. The best results we obtained were accuracy of 94.7%, recall of 94.8%, F1 score of 94.8%, precision of 95.3%, specificity of 94.5% and AUC of 0.98.

In order to more intuitively show whether our proposed method has the ability to improve lr model and whether the results of the proposed model can compete with the ensemble learning algorithm, which has obvious advantages in classification task. During the experiment, we applied lr model without feature combination effect on the same dataset. Additionally, the prediction results of lighthGBM model on each dataset were output. By comparing the results on each dataset, it can be seen that the best classification results on each dataset are almost produced by the lr model with the feature combination effect of lighthGBM. This showed that our proposed method has the ability to improve the classification performance of lr model. Additionally, our proposed method has a competitive advantage.

In addition, we compared the results obtained by our proposed method with those obtained on Z-AlizadehSani dataset reported in previous literature (See Table 21). As shown in Table 21, our proposed method is very competitive. On the Z-Alizadeh Sani dataset, our proposed method achieves better results than existing research. It is worth noting that there are many cell values labeled as 'N' in Table 21, which means that the corresponding indicators are not reported in the literature. However, these indicators are very valuable for evaluating the performance and stability of medical models, especially recall, F1 score and AUC indicators.

To sum up the above, the main contributions of this paper in technology and experimental design are as follows: (1) in the prediction task of CAD, we applied the automatic feature combination function based on lighthGBM model to improve the performance of lr classifier; (2) five sampling methods were applied to Z-Alizadeh Sani dataset to solve the class imbalance problem of dataset; and (3) recall, F1 score, ROC curve and AUC of metrics, which are critical to the evaluation of medical models, were used to evaluate the models. At the same time, our article also has some limitations, such as: (1) more feature combination methods were not tried; and (2) more classification algorithms were not tried. These are also our future research directions.

**Table 21.** The comparison of classification results between our study and other studies on the Z-Alizadeh Sani dataset.

Method	Accuracy %	Recall %	F1%	Precision %	Specificity %	AUC
SMO [33]	92.09	97.22	N	N	79.31	N
SMO + IG [30]	94.08	96.30	N	N	88.51	N
KNN [43]	90.91	93.33	93.33	93.33	85.71	N
NN + genetic algorithm [32]	93.85	97	N	N	92	N
NB + genetic algorithm [69]	88.16	88.00	N	N	87.78	N
Ensemble [70]	86.49	73.61	0.75	N	91.67	0.83
SVM + feature engineering + 500 examples [37]	96.4	100	N	N	88.1	0.92
NE-nu-SVC [35]	94.66	94.70	94.70	94.70	N	0.966
N2GC-nuSVM [36]	93.08	N	91.51	N	N	N
XGBoost + hybrid FSA + FA + ETCA + SMOTE [31]	92.58	92.99	90.62	92.59	N	N
Hybrid PSO-EmNN coupled with feature selection [38]	88.34	91.85	92.12	92.37	78.98	N
XGBoost + feature construction + SMOTE [48]	94.7	96.1	94.6	93.4	93.2	0.98
GSVMA [47]	89.45	81.22	80.49	N	100	100
C-CADZ [41]	97.37	98.15	N	N	95.45	N
CART [46]	92.41	98.61	N	N	77.01	N
lightGBM + lr + SMOTE_Tomek *	94.7	94.8	94.8	95.3	94.5	0.98

\* Our proposed methods.

## 7. Conclusions

CAD is the major cause of global health burden and death. The early, accurate and rapid diagnosis of CAD is the only effective way to reduce the damage of CAD. In this paper, we proposed a model that can accurately detect CAD according to the results of clinical routine examination in the early stage. The model took the logistic regression algorithm as the base classifier, and combined the feature combination processing based on lighthGBM algorithm and the class balance method of resampling. In addition, according to the characteristics of the dataset, the corresponding data preprocessing method was applied. A 10-fold cross validation technique was used to test the robustness of the model. Accuracy, recall, specificity, precision, F1 score, ROC curve and AUC indicators were used to evaluate the model. The experimental results and analysis show that our proposed model has strong advantages in the models used for CAD detection. Our proposed model can be used to assist doctors in making decisions.

**Author Contributions:** Conceptualization, Y.Y., J.T. and X.W.; methodology, S.Z., Y.Y., J.Y. and Z.Y.; software, S.Z.; validation, S.Z.; formal analysis, S.Z., J.T., X.W., J.Y. and Z.Y.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z. and Y.Y.; supervision, Y.Y., J.T. and X.W.; funding acquisition, J.T. 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 Basic Research Development Program, China (2015CB351900).

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

## Appendix A

**Table A1.** A list of abbreviations used in this paper.

Abb	Full Name	Abb	Full Name
CAD	coronary artery disease	ADASYN	adaptive synthetic
DALYs	disability-adjusted life years	SVM	support vector machines



**Table A1.** *Cont.*

Abb	Full Name	Abb	Full Name
CVD	cardiovascular disease	KNN	k-nearest neighbors
IHD	ischemic heart disease	PSO	particle swarm optimization
HIV	human immunodeficiency virus	PCA	principal component analysis
AIDS	acquired immunodeficiency syndrome	RF	random forest
LAD	left anterior descending coronary artery	ET	extreme tree
LCX	left circumflex coronary artery	LDA	linear discriminant analysis
RCA	right coronary artery	GR	gain ratio
ECG	electrocardiogram	IG	information gain
FFR	fractional flow reserve	FLDA	fisher linear discriminant analysis
IVUS	intravenous ultrasound	CS	chi-square
COPD	chronic obstructive pulmonary disease	CART	classification and regression tree
SMO	sequence minimum optimization	ROC	receiver operating characteristic
SMOTE	synthetic minority oversampling technology	AUC	area under curve

Abb, Abbreviations.

**Table A2.** The hyperparameters setting of the proposed model on the dataset processed by data standardization and SMOTE\_Tomek method.

Model	Algorithms	Hyperparameters	Value of Hyperparameters
lightGBM + lr	lightGBM	boosting_type	'gbdt'
		objective	'binary'
		learning_rate	0.52488778
		n_estimators	231
		max_depth	2
		num_leaves	3
		max_bin	74
		min_data_in_leaf	21
		bagging_fraction	0.676450
		bagging_freq	2
		feature_fraction	0.610
		C	1.79532
		class_weight	{0:0.5,1:0.5}
		dual	False
fit_intercept	True		
intercept_scaling	1		
max_iter	4		
multi_class	'ovr'		
n_jobs	1		
penalty	'l1'		
random_state	69		
solver	"liblinear"		
tol	0.01		
verbose	0		
warm_start	False		

## References

1. World Health Organization. *World Health Statistics 2019: Monitoring Health for the SDGs, Sustainable Development Goals*; World Health Organization: Geneva, Switzerland, 2019.
2. 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](#)]
3. Roth, G.A.; Mensah, G.A.; Fuster, V. The Global Burden of Cardiovascular Diseases and Risks: A Compass for Global Action. *J. Am. Coll. Cardiol.* **2020**, *76*, 2980–2981. [[CrossRef](#)] [[PubMed](#)]
4. 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](#)]
5. Roth, G.A.; Mensah, G.A.; Johnson, C.O.; Addolorato, G.; Ammirati, E.; Baddour, L.M.; Barengo, N.C.; Beaton, A.Z.; Benjamin, E.J.; Benziger, C.P.; et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *J. Am. Coll. Cardiol.* **2020**, *76*, 2982–3021. [[CrossRef](#)] [[PubMed](#)]
6. GBD. 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](#)]
7. 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.
8. Jayaraman, V.; Sultana, H.P. Artificial 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](#)]
9. 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.
10. 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](#)]
11. Acharya, U.R.; Fujita, H.; Oh, S.L.; Hagiwara, Y.; Tan, J.H.; Adam, M. Application of Deep Convolutional Neural Network for Automated Detection of Myocardial Infarction Using ECG Signals. *Inf. Sci.* **2017**, *415*, 190–198. [[CrossRef](#)]
12. Gupta, V.; Mittal, M. Arrhythmia Detection in ECG Signal Using Fractional Wavelet Transform with Principal Component Analysis. *J. Inst. Eng. India Ser. B* **2020**, *101*, 451–461. [[CrossRef](#)]
13. Liu, W.; Huang, Q.; Chang, S.; Wang, H.; He, J. Multiple-Feature-Branch Convolutional Neural Network for Myocardial Infarction Diagnosis Using Electrocardiogram. *Biomed. Signal Process. Control* **2018**, *45*, 22–32. [[CrossRef](#)]
14. Tan, J.H.; Hagiwara, Y.; Pang, W.; Lim, I.; Oh, S.L.; Adam, M.; Tan, R.S.; Chen, M.; Acharya, U.R. Application of Stacked Convolutional and Long Short-Term Memory Network for Accurate Identification of CAD ECG Signals. *Comput. Biol. Med.* **2018**, *94*, 19–26. [[CrossRef](#)] [[PubMed](#)]
15. Acharya, U.R.; Fujita, H.; Oh, S.L.; Adam, M.; Tan, J.H.; Chua, K.C. Automated Detection of Coronary Artery Disease Using Different Durations of ECG Segments with Convolutional Neural Network. *Knowl. Base Syst.* **2017**, *132*, 62–71. [[CrossRef](#)]
16. Zihlmann, M.; Perekrestenko, D.; Tschannen, M. Convolutional Recurrent Neural Networks for Electrocardiogram Classification. *arXiv* **2017**, arXiv:1710.06122v1.
17. Gupta, V.; Mittal, M.; Mittal, V. A Novel FrWT Based Arrhythmia Detection in ECG Signal Using YWARA and PCA. *Wirel. Pers. Commun.* **2021**, *1*, 1–18. [[CrossRef](#)]
18. Gupta, V.; Mittal, M.; Mittal, V.; Gupta, A. An Efficient AR Modelling-Based Electrocardiogram Signal Analysis for Health Informatics. *Int. J. Med. Eng. Inform.* **2022**, *14*, 74–89.
19. Patidar, S.; Pachori, R.B.; Acharya, U.R. Automated Diagnosis of Coronary Artery Disease Using Tunable-Q Wavelet Transform Applied on Heart Rate Signals. *Knowl.-Based Syst.* **2015**, *82*, 1–10. [[CrossRef](#)]
20. Sridhar, C.; Acharya, U.R.; Fujita, H.; Bairy, G.M. Automated Diagnosis of Coronary Artery Disease using Nonlinear Features Extracted from ECG Signals. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016; pp. 545–549. [[CrossRef](#)]
21. Altan, G.; Allahverdi, N.; Kutlu, Y. Diagnosis of Coronary Artery Disease Using Deep Belief Networks. *Eur. J. Eng. Nat. Sci.* **2017**, *2*, 29–36.
22. Sharma, M.; Acharya, U.R. A New Method to Identify Coronary Artery Disease with ECG Signals and Time-Frequency Concentrated Antisymmetric Biorthogonal Wavelet Filter Bank. *Pattern Recognit. Lett.* **2019**, *125*, 235–240. [[CrossRef](#)]
23. Zreik, M.; van Hamersvelt, R.W.; Wolterink, J.M.; Leiner, T.; Viergever, M.A.; Išgum, I. A Recurrent CNN for Automatic Detection and Classification of Coronary Artery Plaque and Stenosis in Coronary CT Angiography. *IEEE Trans. Med. Imaging* **2018**, *38*, 1588–1598. [[CrossRef](#)]
24. Sprem, J.; de Vos, B.D.; de Jong, P.A.; Viergever, M.A.; Išgum, I. Classification of Coronary Artery Calcifications According to Motion Artifacts in Chest CT Using a Convolutional Neural Network. In Proceedings of the Medical Imaging 2017: Image Processing, Orlando, FL, USA, 11–16 February 2017; International Society for Optics and Photonics: Orlando, FL, USA, 2017; Volume 10133.
25. Kim, G.Y.; Lee, J.H.; Hwang, Y.N.; Kim, S.M. A Novel Intensity-Based Multi-Level Classification Approach for Coronary Plaque Characterization in Intravascular Ultrasound Images. *Biomed. Eng. Online* **2018**, *17*, 200–213. [[CrossRef](#)]

26. Gessert, N.; Lutz, M.; Heyder, M.; Latus, S.; Leistner, D.M.; Abdelwahed, Y.S.; Schlaefer, A. Automatic Plaque Detection in IVOCT Pullbacks Using Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2018**, *38*, 426–434. [[CrossRef](#)] [[PubMed](#)]
27. Yang, J.L.; Zhang, B.; Wang, H.R.; Lin, F.; Han, Y.C.; Lin, X.L. Automated Characterization and Classification of Coronary Atherosclerotic Plaques for Intravascular Optical Coherence Tomography. *Biocybern. Biomed. Eng.* **2019**, *39*, 719–727. [[CrossRef](#)]
28. Martin-Isla, C.; Campello, V.M.; Izquierdo, C.; Raisi-Estabragh, Z.; Baessler, B.; Petersen, S.E.; Lekadir, K. Image-Based Cardiac Diagnosis With Machine Learning: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 1. [[CrossRef](#)] [[PubMed](#)]
29. Chen, C.; Qin, C.; Qiu, H.Q.; Tarroni, G.; Duan, J.M.; Bai, W.J.; Rueckert, D. Deep Learning for Cardiac Image Segmentation: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 25. [[CrossRef](#)] [[PubMed](#)]
30. 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](#)] [[PubMed](#)]
31. Nasarian, E.; Abdar, M.; Fahami, M.A.; Alizadehsani, R.; Hussain, 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](#)]
32. Arabasadi, Z.; Alizadehsani, R.; Roshanzamir, M.; Moosaei, H.; Yarifard, A.A. Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Net-Work-Genetic Algorithm. *Comput. Methods Programs Biomed.* **2017**, *141*, 19–26. [[CrossRef](#)]
33. 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.
34. 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](#)]
35. 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](#)]
36. 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](#)]
37. 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](#)] [[PubMed](#)]
38. 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](#)]
39. Wang, J.K.; Liu, C.C.; Li, L.P.; Li, W.; Yao, L.K.; Li, H.; Zhang, H. A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease. *IEEE Access* **2020**, *8*, 37124–37133. [[CrossRef](#)]
40. Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. *BioMed Res. Int.* **2020**, *2020*, 9816142. [[CrossRef](#)] [[PubMed](#)]
41. Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. C-CADZ: Computational Intelligence System for Coronary Artery Disease Detection Using Z-Alizadeh Sani Dataset. *Appl. Intell.* **2021**, *52*, 2436–2464. [[CrossRef](#)]
42. 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.
43. 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](#)]
44. Alizadehsani, R.; Habibi, J.; Alizadehsani, Z.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B. Diagnosis of Coronary Artery Disease Using Data Mining Based on Lab Data and Echo Features. *J. Med. Bioeng.* **2012**, *1*, 26–29. [[CrossRef](#)]
45. Yadav, C.; Lade, S.; Suman, M.K. Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining. *Int. J. Comput. Appl.* **2014**, *87*, 9–13. [[CrossRef](#)]
46. Ghiasi, M.M.; Zendejboudi, S.; Mohsenipour, A.A. Decision Tree-Based Diagnosis of Coronary Artery Disease: CART Model. *Comput. Methods Programs Biomed.* **2020**, *192*, 105400. [[CrossRef](#)]
47. Joloudari, J.H.; Azizi, F.; Nematollahi, M.A.; Alizadehsani, R.; Hassannatajjeloudari, E.; Nodehi, I.; Mosavi, A. GSVMA: A Genetic Support Vector Machine ANOVA Method for CAD Diagnosis. *Front. Cardiovasc. Med.* **2022**, *8*, 760178. [[CrossRef](#)]
48. 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. [[CrossRef](#)]
49. 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](#)] [[PubMed](#)]
50. 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, Vancouver, BC, Canada, 1 December 2004; pp. 825–832.

51. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
52. Xu, X.L.; Chen, W.; Sun, Y.F. Over-Sampling Algorithm for Imbalanced Data Classification. *J. Syst. Eng. Electron.* **2019**, *30*, 1182–1191. [[CrossRef](#)]
53. 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.
54. 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.
55. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Lect. Notes Artif. Intell.* **2005**, *3644*, 878–887.
56. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline Over-sampling for Imbalanced Data Classification. In Proceedings of the Fifth International Workshop on Computational Intelligence & Applications, Hiroshima University, Hiroshima, Japan, 10–12 November 2009.
57. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [[CrossRef](#)]
58. Wang, M.Y.; Wei, Z.H.; Jia, M.; Chen, L.Z.; Ji, H. Deep Learning Model for Multi-Classification of Infectious Diseases from Unstructured Electronic Medical Records. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 41. [[CrossRef](#)] [[PubMed](#)]
59. Santos, L.I.; Camargos, M.O.; Silveira, M.F.; D’Angelo, M.F.S.V.; Mendes, J.B.; Medeiros, E.E.C.d.; Guimarães, A.L.S.; Palhares, R.M. Decision Tree and Artificial Immune Systems for Stroke Prediction in Imbalanced Data. *Expert Syst. Appl.* **2022**, *191*, 116221. [[CrossRef](#)]
60. Sapra, V.; Saini, M.L.; Verma, L. Identification of Coronary Artery Disease using Artificial Neural Network and Case-Based Reasoning. *Recent Adv. Comput. Sci. Commun.* **2021**, *14*, 2651–2661. [[CrossRef](#)]
61. Chen, X.; Fu, Y.; Lin, J.; Ji, Y.; Fang, Y.; Wu, J. Coronary Artery Disease Detection by Machine Learning with Coronary Bifurcation Features. *Appl. Sci.* **2020**, *10*, 7656. [[CrossRef](#)]
62. Guo, H.; Tang, R.; Ye, Y.; Li, Z.; He, X. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 19–25 August 2017; pp. 1725–1731.
63. Ke, G.L.; Meng, Q.; Finley, T.; Wang, T.F.; Chen, W.; Ma, W.D.; Ye, Q.W.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 1–9.
64. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
65. 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](#)] [[PubMed](#)]
66. 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.
67. Tian, X.; Wang, J.; Wen, Y.; Ma, H. Multi-Attribute Scientific Documents Retrieval and Ranking Model Based on GBDT and LR. *Math. Biosci. Eng.* **2022**, *19*, 3748–3766. [[CrossRef](#)] [[PubMed](#)]
68. Zabor, E.C.; Reddy, C.A.; Tendulkar, R.D.; Patil, S. Logistic Regression in Clinical Studies. *Int. J. Radiat. Oncol. Biol. Phys.* **2021**, *112*, 271–277. [[CrossRef](#)] [[PubMed](#)]
69. 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.
70. Cüvitoğlu, A.; Işık, Z. Classification of Cad Dataset by Using Principal Component Analysis and Machine Learning Approaches. In Proceedings of the 5th International Conference on Electrical and Electronic Engineering (ICEEE), Istanbul, Turkey, 3–5 May 2018; pp. 340–343.