


Article

A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm

Jian Yang *  and Jinhan Guan

School of Information, Shanxi University of Finance and Economics, Taiyuan 030006, China

* Correspondence: yangj@sxufe.edu.cn

Abstract: In today's world, heart disease is the leading cause of death globally. Researchers have proposed various methods aimed at improving the accuracy and efficiency of the clinical diagnosis of heart disease. Auxiliary diagnostic systems based on machine learning are designed to learn and predict the disease status of patients from a large amount of pathological data. Practice has proved that such a system has the potential to save more lives. Therefore, this paper proposes a new framework for predicting heart disease using the smote-xgboost algorithm. First, we propose a feature selection method based on information gain, which aims to extract key features from the dataset and prevent model overfitting. Second, we use the Smote-Enn algorithm to process unbalanced data, and obtain sample data with roughly the same positive and negative categories. Finally, we test the prediction effect of Xgboost algorithm and five other baseline algorithms on sample data. The results show that our proposed method achieves the best performance in the five indicators of accuracy, precision, recall, F1-score and AUC, and the framework proposed in this paper has significant advantages in heart disease prediction.

Keywords: feature selection; smote-xgboost; heart disease prediction



Citation: Yang, J.; Guan, J. A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm. *Information* **2022**, *13*, 475. <https://doi.org/10.3390/info13100475>

Academic Editor: Anirban Bandyopadhyay

Received: 20 July 2022

Accepted: 27 September 2022

Published: 2 October 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

Cardiovascular disease, which is mostly caused by heart disease, causes over one-third of all annual deaths globally [1]. Many academics have developed heart disease diagnostic systems by utilizing machine learning to obtain useful information from existing medical databases in an effort to change the current situation. Using these diagnostic tools can assist clinical decisions on a medical diagnosis, speed up the diagnosis process, and uncover disease-related knowledge that will save the lives of more people.

When making a disease diagnosis, there is a great deal of information on the patient's pathology, which is expressed in the dataset as a good amount of features [2]. Each feature has a unique effect on the disease diagnosis results. Often, a few major features contribute to the diagnosis of the presence or absence of disease. Applying some feature selection methods before training the model can help with selection of some key features, thus leading to good prediction results in a shorter period of time. Most disease datasets have an imbalanced distribution, with more samples falling into the negative category and fewer samples falling into the positive category. The distribution of the dataset can be modified using specific data processing techniques, which will rebalance it and boost the validity of the model.

Machine learning algorithms had a significant advantage in dealing with problems with complex and nonlinear features. A number of diseases classification and prediction challenges, including early warning for Electrocardiogram (ECG) detection [3] and prediction related to congenital heart disease [4], have been successfully addressed by the use of several algorithms, including LR, SVM, and KNN, etc.

Ensemble learning [5] is the basis for the construction of many machine learning algorithms. The fundamental idea is to combine the advantages of several poor classifiers

to produce a model with superior overall performance. Bagging and boosting are the two primary model combination techniques; bagging integrates multiple underfitting weak classifiers, and boosting integrates multiple overfitting weak classifiers. Xgboost is an efficient implementation of ensemble learning, whose main idea is boosting and introducing regular terms in the objective function to prevent overfitting.

We propose a heart disease prediction model by employing the smote-xgboost algorithm. The model was trained using real pathological data from cardiac patients. Among them, Major Adverse Cardiovascular Events (MACCE) are the prediction target, and the occurrence of MACCE is a key indicator to evaluate the success of coronary heart disease surgery. In summary, we make important contributions as follows.

- To remove the crucial features from the dataset, an information gain-based feature selection method is used.
- Use a technique that combines undersampling and oversampling to handle uneven data on the selected dataset.
- Using the preprocessed dataset, validate efficacy of xgboost. Additionally, assess the ability of the xgboost algorithm with five baseline methods using a confusion matrix.

The remaining portions of the essay are structured as follows. Section 2 summarizes the most recent research on heart disease prediction. Section 3 is a brief description of the dataset and an introduction to the algorithms applied in the framework. Section 4 is a statistical description of the dataset and a comparison and evaluation of the experimental models, and Section 5 provides a conclusion and outlook.

2. Related Work

One of the major applications of machine learning in recent years has been the prediction of heart disease, which has had some success. Some scholars have concentrated on the innovation of data processing techniques such as feature selection, and some scholars have focused on innovation from the perspective of prediction algorithms.

Modepalli et al. [6] utilized a new model (DT + RF) to predict the occurrence (or non-occurrence) of heart disease. They chose the UCI dataset to validate the reliability of the hybrid model, comparing the prediction outcomes of the hybrid model and any single algorithm in the hybrid model, respectively. It is found that the hybrid model has a significant advantage over the single algorithm in terms of performance in the evaluation metric of accuracy, with a 7% to 9% improvement.

Joo et al. [7] used a dataset of cardiovascular disease with the same features but different years of return visits to train the model. The authors selected 25 features from the dataset by combining health examination results and questionnaire responses, and used four machine learning models to predict the 2-year and 10-year cardiovascular disease risk, respectively. In particular, they found that the accuracy of each model improved somewhat if physician medication information was taken into account when performing feature selection, and that medication information had a strong effect on the prediction of short-term data in this study.

Li et al. [8] put out a feature selection approach fast conditional mutual information (FCMIM) based on conditional mutual information. They employed four common feature selection algorithms and FCMIM on the Cleveland dataset and used six machine learning algorithms to train the model. The results suggested the use of this novelty feature selection method, with the highest accuracy of 92.37% for the combination of FCMIM and SVM.

Ali et al. [9] used a feature fusion technique to process low-dimensional data extracted from medical records and sensor data. Then, they employed a feature selection strategy relating to information gain and feature ranking to obtain the dataset. They achieved prediction accuracy of 98.5% by applying an ensemble deep learning algorithm.

Rahim et al. [10] applied an oversampling technique to balance the data, and also used the mean value method to fill in the missing values and feature importance method for feature selection. They selected three datasets (including the Framingham dataset and the Cleveland dataset). After data preprocessing on each of the three datasets, the predictive

effectiveness of the new ensemble model (KNN and LR) with and without feature selection was compared. The results fully validated the advantages of the new ensemble model, in which the accuracy of the new model with feature selection was as high as 99.1%.

Ishaq et al. [11] used the feature importance of random forest to rank the features and select the features with higher scores, and also employed the SMOTE technique to balance the data. They compared the prediction performance of nine commonly used algorithms on data treated with SMOTE and on unbalanced data without treatment, where it was found that the prediction accuracy of each model was significantly improved on balanced data.

Khurana et al. [12] found that SVM outperformed all other machine learning algorithms when testing their results on the Cleveland dataset by applying five feature selection techniques. The prediction accuracy of each machine learning algorithm improved to a different extent after applying the feature selection methods, where the feature selection methods with Chi-Square and information gain were applied. The accuracy of the combination of Chi-Square and information gain and SVM both reached 83.41%.

Ashri et al. [13] applied a genetic-algorithm-based feature selection Simple Genetic Algorithm (SGA) and trained model by using UCI dataset. Two algorithms with the highest accuracy were selected to propose a hybrid ensemble learning model based on decision trees and random forests, and found that the accuracy of the ensemble learning model reached 98.18%.

Bashir et al. [14] proposed a new ensemble learning combinatorial voting approach, in which four datasets were selected from the UCI database to validate six machine learning algorithms and five ensemble models with a combination of these six algorithms. They found that the accuracy of the ensemble models was generally greater compared to the individual algorithms, in which the average accuracy of the five ensemble models reached 83%. The proposed combination can be extended to bagging and boosting to further improve the accuracy.

In conclusion, data preprocessing, such as data standardization and feature selection, can effectively raise the value of the dataset and greatly enhance the accuracy of a model. Additionally, ensemble learning models perform well when dealing with heart disease. The main point of this study is to employ the ensemble learning algorithm Xgboost on a heart disease dataset after performing feature selection and imbalance processing. Finally, by contrasting xgboost with other standard algorithms, the effectiveness and accuracy of the suggested framework in predicting heart disease are confirmed.

3. Method

Figure 1 shows the heart disease prediction framework proposed in this paper.

3.1. Dataset

This paper uses the return visit data of real patients in a hospital as the research sample. We named this the Heart Disease Dataset (HDD). The dataset has a total of 4232 samples and 37 features, including numeric and categories. The predictive target is major adverse cardiovascular and cerebrovascular events (MACCE), where zero indicates no occurrence and one indicates occurrence.

3.2. Data Preprocessing

Data processing is a vital stage before training, since the quality of the data will directly affect the predictions made by the model. We use the following approach to handle missing values. For class variables, we create a new class to represent the null values; for numeric variables, we eliminate the feature columns with missing values rates greater than 70%, citing them as invalid, and replace the remaining feature columns with missing values with the mean values. We also normalize the data using the maximum–minimum norm method to enhance the data’s relevance. The formula is as follows.

$$H = \frac{H^0 - H^{min}}{H^{max} - H^{min}} \times (NH_{max} - NH_{min}) + NH_{min} \quad (1)$$

where Equation (2) denotes the information entropy of feature X , Equation (3) denotes the information entropy of prediction column Y when feature X is known, and Equation (4) denotes the information gain, and the information gain of feature X is the difference between the information entropy of prediction column Y and the conditional entropy of both. Different information gain values are taken for various features in the dataset, and these values are sorted. The features with gains larger than the threshold are regarded as essential features that should be selected. The following is its pseudo code.

After the above preprocessing and feature selection, we get a total of 3527 sample data, as well as 15 features and 1 predicted label. The following Table 1 provides a description of the preprocessed HDD.

Table 1. Description of Features.

Index	Feature	Type	Description
1	Sex	category	Man = 1; Female = 0
2	Stable_CAD	category	Stable CAD = 0; Unstable CAD = 1
3	Age	numeric	Age in years, [20, 86]
4	CVD_history	category	Ischemic cerebrovascular disease = 0; Hemorrhagic cerebral vascular diseases = 1
5	Smoke	category	No smoking history = 0; Have smoking history = 1
6	nitrate	category	Hospitalization without nitrate = 0; Hospitalization with nitrate = 1
7	LVEF	numeric	Left ventricular ejection fraction, [18, 88]
8	HBG	numeric	Hemoglobin, [55, 193.2]
9	BUN	numeric	Blood urea nitrogen, [0.7, 119.0]
10	TC	numeric	Total cholesterol, [73, 589]
11	SCV_number	numeric	SCV_number, [0, 3]
12	DM	category	No diabetes mellitus = 0; Having diabetes mellitus = 1
13	REV_type	category	PCI = 1; CABG = 2
14	LM_lesion	category	No LM_lesion = 0; Having LM_lesion = 1
15	ASA	category	Hospitalization without ASA = 0; Hospitalization with ASA = 1
16	MACCE	category	No MACCE = 1; Occurrence of MACCE = 1

3.4. Imbalance Data Processing Based on Smote-Enn

Due to the low prevalence of most diseases, the distribution of medical datasets is typically imbalanced, exhibiting significant differences in the number of samples from various categories in the dataset. When the model is trained with imbalanced datasets, the performance and dependability of the model are decreased. Table 2 shows the category distribution of the target MACCE of the unbalanced HDD employed in this paper, where the ratio between 0 (not occurring) and 1 (occurring) reached 9:1.

Table 2. The distribution of MACCE in HDD.

MACCE	0	1	Total
Number	3204	323	3527
Percentage	90.84%	9.16%	100%

To obtain balanced data, there are three basic strategies: (1) expanding the sample size from the minority class (oversampling); (2) decreasing the number of samples from the majority class (undersampling); and (3) combining undersampling and oversampling. The undersampling method removes samples from the majority class at random, which may lead to a loss of crucial information that has a considerable impact on the learning task. The oversampling method directly resamples samples from the minority class, which may result in overfitting of the model. Furthermore, several researchers have shown that mixed methods are superior to single methods when processing datasets [16,17].

In this research, a hybrid technique called SMOTE-ENN [18] is utilized to handle imbalanced data. SMOTE is an oversampling algorithm that employs a method of interpolating samples from the minority class. By removing samples that do not fall into the categories that account for the majority of the k -nearest neighbor samples, the ENN algorithm, which is an undersampling algorithm, decreases the amount of samples from the majority class. In this paper, the SMOTE algorithm is used to undersample the category of MACCE of one until the balance between the samples in the majority and minority groups is reached. Then, the ENN algorithm is applied to remove the overlapping samples in each of the two categories until the dataset is rebalanced. Using this hybrid technique, the minority class of HDD has a proportion of 61.67 percent, increasing from 9.16 percent. In Algorithm 2, the SMOTE-ENN pseudocode is included.

Algorithm 2: The pseudo code of Smote-Enn.

Input : Heart Disease Dataset HDD;

Process:

```

1 foreach sample  $s_i$  in the minority class of HDD do
2   Calculate the K-nearest neighbor samples  $ks_i$  of  $s_i$ ;
3   Construct a new data sample  $ns = s_i + (\hat{s}_i - s_i) + \delta$ ;
4   Add the generated sample  $ns$  to HDD;
5 foreach sample  $h_i$  in HDD do
6   if  $h_i$  class  $\neq$  majority class of  $k$ -nearest neighbors then
7     Remove  $h_i$ ;

```

Output: Balanced dataset HDD

3.5. XGBoost

Xgboost is an implementation of the ensemble learning algorithm boosting [19]. The fundamental principle of the Xgboost is to train the model using residuals. The outcome of the most recent tree training is utilized as the input for the subsequent iteration, and the error is progressively decreased over numerous serial iterations. Finally, all weak learners are linearly weighted to produce the ensemble learner.

Additionally, when training the Xgboost tree, the effective splitting point is chosen using an information-gain-based greedy algorithm. To better optimize the objective function, Xgboost uses a second-order Taylor expansion to approximate the objective function, and the optimal solution is the quadratic optimal solution. Furthermore, a regular term is added to regulate the spanning tree's complexity, lowering the possibility of overfitting the model. The loss function is as follows

$$f_{obj}^{(t)} = \sum_n^i [L(y_i, \hat{y}_i^{t-1}) + f_t(x_i)] + \frac{1}{2} L''(y_i, \hat{y}_i^{t-1}) + f_t^2(x_i) + \Omega(f_t) \quad (5)$$

$$\Omega(f_t) = \frac{1}{2} \lambda \sum_j^T \|W_j\|_2 + \gamma T \quad (6)$$

W_j stands for the leaf node weights, T stands for the total number of nodes, and λ and γ are hyperparameters that control the node complexity.

The Xgboost technique utilizes the shrinkage strategy [20] to ensemble weak learners and decrease the likelihood of overfitting the model. This ensemble takes the form shown below.

$$F_m(X) = F_{m-1}(X) + \eta f_m(X), 0 < \eta < 1 \quad (7)$$

where $f_m(X)$ denotes the m th iteration to generate the weak learner and $F_m(X)$ denotes the m th iteration to generate the integrated learner. Since the parameter η has a strong negative correlation with the number of iterations, the model often has better generalization properties when η has a smaller value.

Moreover, Xgboost adopts the Parzen estimation tree strategy to automatically optimize the hyperparameters in the model for optimal prediction, as well as the block technique to enhance the capability of the model to handle large amounts of data and improve its training efficiency.

3.6. Baseline Algorithms

Five machine learning methods are chosen as the baseline algorithms in this paper. This research compares the prediction performance of the baseline algorithm with Xgboost to illustrate the utility of Xgboost in predicting heart disease. The following is an overview of the baseline algorithms.

3.6.1. Random Forest

RF [21] is an ordinary bagging algorithm. Unlike conventional decision trees, RF trains each classifier using a randomly chosen subset of the dataset and a randomly chosen subset of the features. Each trained classifier produces different prediction results for the same input. Voting for the output of each trained classifier, typically using the plurality or the mean, leads to the final prediction result. As the features of the algorithm are randomly divided, it will increase the diversity of its classifiers and thus enhance the model's capacity for generalization.

3.6.2. K-Nearest Neighbor

KNN [22] is a form of lazy learning in which KNN learns after receiving the test samples, and the time overhead of the algorithm training samples is zero. The algorithm in the test sample will utilize the distance as the metric to find the k sample points that are closest to each test sample point, and it will use the category information of the k sample points as the judgment basis. The category with the greatest percentage of the k sample points is typically utilized as the test sample in the binary classification issue.

3.6.3. Logistic Regression

LR [23] is a variant of the linear regression algorithm. For the binary classification issue, Logistic Regression applies a logistic function to convert values predicted by a linear regression technique into discrete values (i.e., zero and one) if the predicted value is larger than zero, then one otherwise. Below is a diagram of the logistic function.

$$p = \frac{1}{1 + e^{-y}} \quad (8)$$

3.6.4. Decision Tree

DT [24] is a widely used classification algorithm, which can be categorized into three types according to the varied methods of generating trees. These categories include the decision tree based on information gain, which represents the ID3 tree, the decision tree algorithm based on gain rate, which represents the C4.5 tree, and the decision tree based on the Gini index, which represents the CART tree. The decision tree algorithm will also employ prepruning and postpruning procedures to prevent overfitting and enhance the system's capacity for generalization.

3.6.5. Naïve Bayes

NB [25] is a classification algorithm based on event probability and misclassification loss. The main advantage of NB is that it adopts the attribute conditional independence assumption strategy to avoid the combinatorial explosion problem that occurs when

computing posterior probabilities. According to the attribute conditional independence assumption, the class conditional probabilities are recast as follows.

$$P(C|x) = \frac{P(C)}{p(x)} \prod_{i=1}^d P(x_i|C) \quad (9)$$

The test results are then categorized in accordance with the corresponding probability.

$$C_{nb} = \operatorname{argmax}_C P(C) \prod_{i=1}^d P(x_i|C) \quad (10)$$

4. Performance Evaluation

4.1. Result of Exploratory Data Analysis

Exploratory data analytics were carried out on this dataset to better understand its characteristics. The following subsection provides a description of the analyses' observations.

The frequency distribution histogram provides a rapid overview of the data's dispersion and central tendency. The distribution of various features is visually represented by the height of each rectangle in Figure 2, which shows the frequency of occurrence of the values. Additionally, the ability of the model to predict outcomes is impacted by the degree of feature correlation.

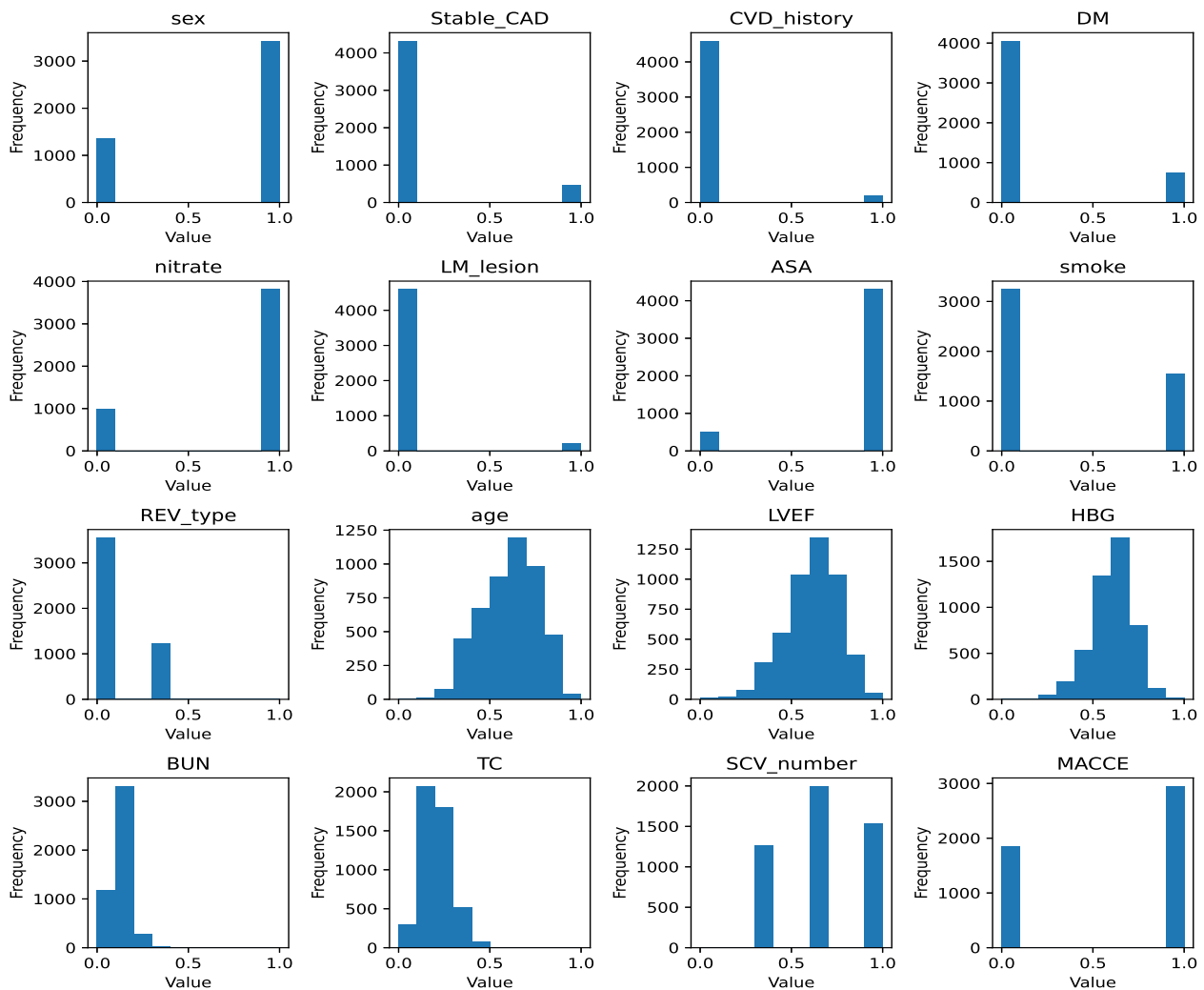


Figure 2. The frequency distribution histogram of HDD.

In this paper, we utilize Pearson correlation coefficients to calculate the correlation coefficients between features and a heat map to show the level of correlation between features. Each row and each column in Figure 3 represent the correlation coefficient between the corresponding features. It can be inferred that the chosen features have independent effects on the prediction column MACCE, since each feature's correlation coefficient in the figure is less than 0.5.

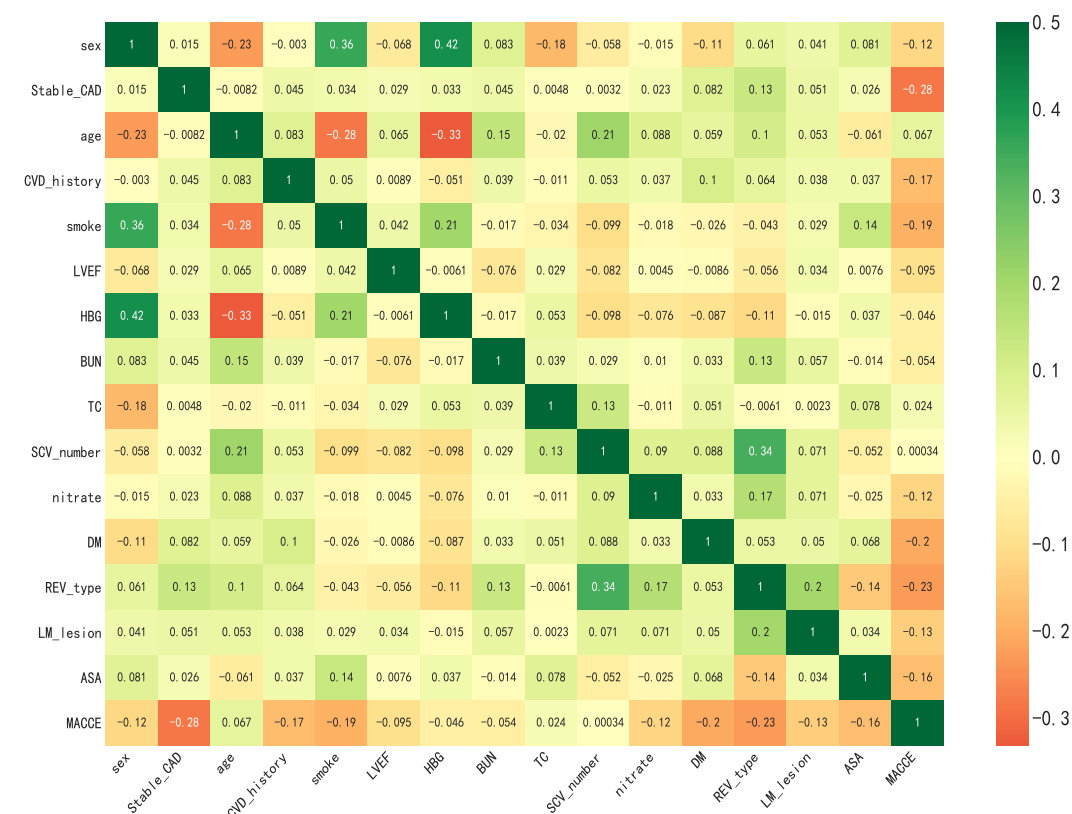


Figure 3. The correlation matrix of features.

4.2. Cross Validation

In this paper, the training and test sets are produced using the five-fold cross-validation approach. The dataset is initially sampled in layers to produce 5 subsets (D1–D5) that are mutually exclusive, equal in size, and have a dependable distribution. We use one subset as the test set and the remaining subsets as the training set in each round. The average of the five sets of results are then used to get the final result. Figure 4 shows a schematic of the five-fold cross-validation approach.

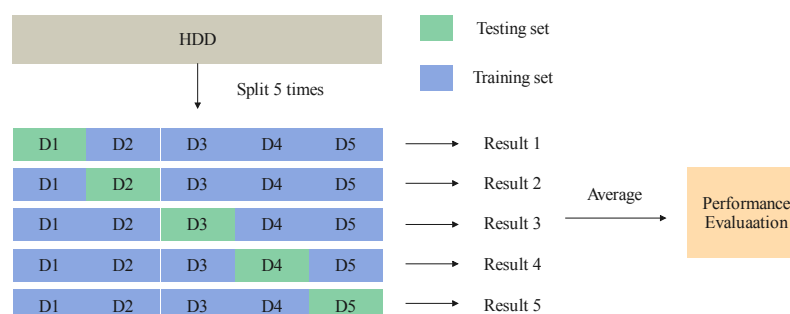


Figure 4. Graphical representation of 5-fold cross-validation.

4.3. Performance Measure

The prediction performance of the algorithm is evaluated in this research using five performance measures based on the confusion matrix. Figure 5 illustrates the binary classification problem's confusion matrix structure. Distinct predicted and true values can be merged into four cases: TP, TN, FP, and FN.

		Predicted class	
		0(Negative)	1(Positive)
Actual class	0(Negative)	TN(True Negative)	FP(False Positive)
	1(Positive)	FN(False Negative)	TP(True Positive)

Figure 5. Confusion matrix.

Using the data from the confusion matrix, the four evaluation indicators can be calculated using the formula below.

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

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1 - Score = \frac{1}{2} \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (14)$$

The ROC curve is a tool for examining the capacity of an algorithm for generalization. The False Positive Rate (FPR) is its horizontal axis and the True Positive Rate (TPR) is its vertical axis, both of which are calculated as follows.

$$TPR = \frac{TP}{TP + FN} \quad (15)$$

$$FPR = \frac{FP}{TN + FP} \quad (16)$$

Moreover, the Area Under ROC Curve (AUC) reflects how well a model predicts heart disease.

4.4. The Performance of Algorithms

In this section, the six algorithms are trained using a five-fold cross-validation method, and the proposed framework is validated using preprocessed HDD. The confusion matrix for the six algorithms is displayed in Table 3, which shows in detail the percentage of the number of the four cases TP, FP, TN, FN in the prediction results of the six algorithms. Table 4 depicts the average performance of the six different algorithms on five metrics: Accuracy, Precision, Recall, F1-score, and AUC.

Table 3. The confusion matrix of six algorithms.

Algorithm	Confusion matrix	Description																
RF	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 869</td><td>FP = 91</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 4</td><td>TP = 94</td><td>98</td></tr><tr><td></td><td>873</td><td>185</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 869	FP = 91	960	Actual True	FN = 4	TP = 94	98		873	185		<p>TN: MACCE was correctly predicted not to occur for 869 samples, and the actual sample MACCE does not occur.</p> <p>TP: MACCE was correctly predicted to occur for 94 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 91 samples, and the actual sample MACCE does not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for 4 samples, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 869	FP = 91	960														
Actual True	FN = 4	TP = 94	98															
	873	185																
KNN	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 873</td><td>FP = 87</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 1</td><td>TP = 97</td><td>98</td></tr><tr><td></td><td>874</td><td>176</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 873	FP = 87	960	Actual True	FN = 1	TP = 97	98		874	176		<p>TN: MACCE was correctly predicted not to occur for 873 samples, and the actual sample MACCE did not occur.</p> <p>TP: MACCE was correctly predicted to occur for 97 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 87 samples, and the actual sample MACCE did not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for one sample, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 873	FP = 87	960														
Actual True	FN = 1	TP = 97	98															
	874	176																
LR	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 707</td><td>FP = 253</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 14</td><td>TP = 84</td><td>98</td></tr><tr><td></td><td>721</td><td>337</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 707	FP = 253	960	Actual True	FN = 14	TP = 84	98		721	337		<p>TN: MACCE was correctly predicted not to occur for 707 samples, and the actual sample MACCE does not occur.</p> <p>TP: MACCE was correctly predicted to occur for 84 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 253 samples, and the actual sample MACCE did not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for 14 samples, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 707	FP = 253	960														
Actual True	FN = 14	TP = 84	98															
	721	337																
DT	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 792</td><td>FP = 168</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 8</td><td>TP = 90</td><td>98</td></tr><tr><td></td><td>800</td><td>258</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 792	FP = 168	960	Actual True	FN = 8	TP = 90	98		800	258		<p>TN: MACCE was correctly predicted not to occur for 792 samples, and the actual sample MACCE did not occur.</p> <p>TP: MACCE was correctly predicted to occur for 90 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 168 samples, and the actual sample MACCE did not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for eight samples, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 792	FP = 168	960														
Actual True	FN = 8	TP = 90	98															
	800	258																
NB	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 712</td><td>FP = 248</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 8</td><td>TP = 90</td><td>98</td></tr><tr><td></td><td>720</td><td>338</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 712	FP = 248	960	Actual True	FN = 8	TP = 90	98		720	338		<p>TN: MACCE was correctly predicted not to occur for 712 samples, and the actual sample MACCE did not occur.</p> <p>TP: MACCE was correctly predicted to occur for 90 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 248 samples, and the actual sample MACCE did not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for eight samples, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 712	FP = 248	960														
Actual True	FN = 8	TP = 90	98															
	720	338																
XGBoost	<table><tr><td></td><td>Predicted False</td><td>Predicted True</td><td></td></tr><tr><td>Actual False</td><td>TN = 899</td><td>FP = 61</td><td>960</td></tr><tr><td>Actual True</td><td>FN = 8</td><td>TP = 90</td><td>98</td></tr><tr><td></td><td>907</td><td>151</td><td></td></tr></table>		Predicted False	Predicted True		Actual False	TN = 899	FP = 61	960	Actual True	FN = 8	TP = 90	98		907	151		<p>TN: MACCE was correctly predicted not to occur for 899 samples, and the actual sample MACCE did not occur.</p> <p>TP: MACCE was correctly predicted to occur for 90 samples, and the actual sample MACCE occurred.</p> <p>FP: MACCE was incorrectly predicted to occur for 61 samples, and the actual sample MACCE did not occur.</p> <p>FN: MACCE was incorrectly predicted not to occur for eight samples, and the actual sample MACCE occurred.</p>
		Predicted False	Predicted True															
	Actual False	TN = 899	FP = 61	960														
Actual True	FN = 8	TP = 90	98															
	907	151																

The most crucial metric for evaluating how well a model predicts is accuracy, of which Xgboost achieves 93.44%. Random Forest and K-Nearest Neighbor achieve 91.15% and 91.77%, respectively. Decision Tree comes in at 83.35%, while Naive Bayes and Logistic Regression fall short at 75.5%. With performance rates of 75.85% and 74.81%, respectively, both Naïve Bayes and Logistic Regression underperformed.

The two measures, Precision and Recall, have a tendency to be inversely correlated; that is, when the precision is high, the recall is typically lower, and vice versa when the recall is high. The most confident samples should be chosen in order to raise the number of correct predictions in MACCE, which will reduce the number of FN and lower the recall. The number of FP will rise when the sample size is increased, resulting in low precision and decreasing the ability to predict the occurrence of MACCE to the greatest extent possible. Owing to the dataset's uniform distribution of positive and negative samples and the model preference that the occurrence of MACCE can be predicted, the Precision of the general model in this experiment is lower than its Recall. The weakest performers were Naïve Bayes and logistic regression, earning 71.05% and 70.59%, respectively, while Xgboost had the best F1-Score at 94.86%.

The ROC curves of these algorithms, which represent the AUC metric in the table, are shown in Figure 6. The higher the value of AUC, the stronger the generalization ability of the model, which is expressed in the ROC curve as the curve close to the upper-left corner of the graph. Naïve Bayes had the worst performance, with an AUC value of 70.59%, but Xgboost had the highest AUC value of 92.44%.

When all evaluating metrics are considered, Xgboost algorithm has a significant advantage in predicting the occurrence of MACCE, and K-Nearest Neighbor is the second best. Apart from that, the comparison showed that Naïve Bayes and Logistic Regression both fared poorly.

Table 4. The results of six algorithms.

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9115	0.9026	0.9615	0.9037
KNN	0.9177	0.8878	0.9933	0.9085
Logistic Regression	0.7481	0.7625	0.8645	0.7178
Decision Tree	0.8335	0.8308	0.9197	0.8157
Naïve Bayes	0.7585	0.7486	0.9214	0.7157
XGBoost	0.9344	0.9266	0.9716	0.9486

The 15 features in HDD each have a unique impact on the outcomes of the predictions. Each model prefers different features, and the scores of these features are also varied. The feature importance ranking of Xgboost, RF, LR, and DT is shown in Figure 7 as the KNN and NB algorithms lack an internal evaluation of feature importance.

From the feature ranking in Table 5, it can be seen that Total cholesterol (TC) is an important feature for predicting whether MACCE occurs. Hemoglobin (HBG) and age are two features that appear in the top five important features of all four algorithms, and are also influential factors that cannot be ignored when predicting. In addition, the table illustrates that the ranking of feature importance is similar for Decision Tree and Xgboost, since the two algorithms construct the same tree structure when training.

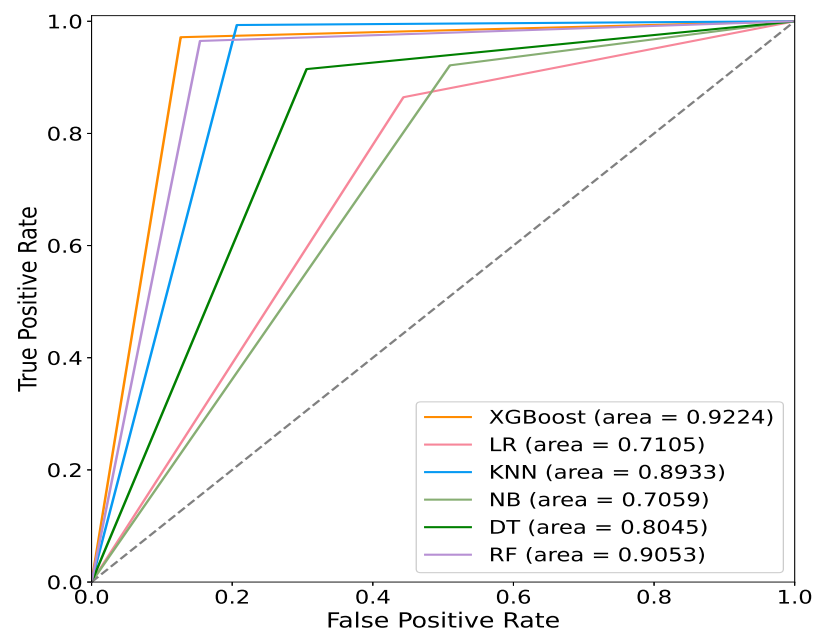


Figure 6. The ROC curve of baseline algorithms and Xgboost.

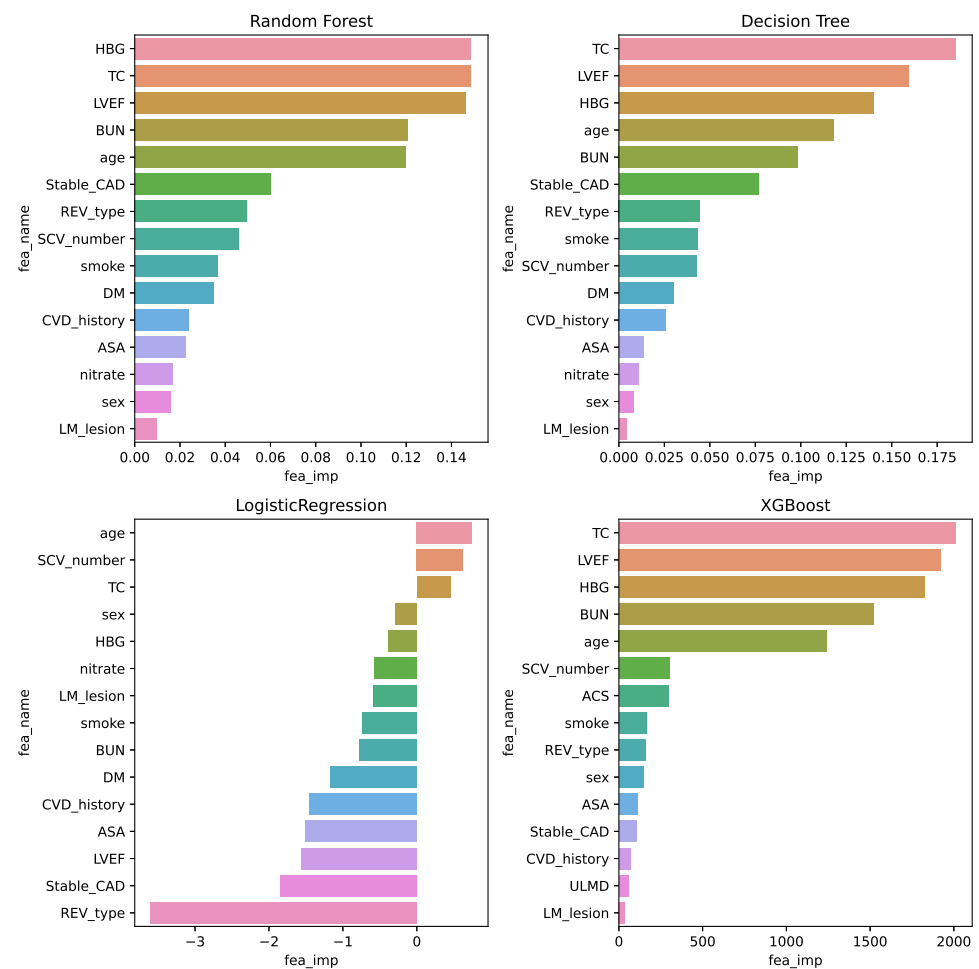


Figure 7. The feature importance of 4 algorithms.

Table 5. Features ranking of 4 algorithms.

Ranking	Random Forest	Decision Tree	Logistic Regression	XGBoost
1	HBG	TC	age	TC
2	TC	LVEF	SCV_number	LVEF
3	LVEF	HBG	TC	HBG
4	BUN	age	sex	BUN
5	age	BUN	HBG	age

In summary, we used multiple evaluation methods to present the prediction results of the Xgboost algorithm and the selected baseline algorithm on the processed dataset. The number of the four prediction outcome cases, TN, TP, FN, and FP, is displayed in the confusion matrix. Accuracy, Precision, Recall, and F1-score were applied based on the confusion matrix, and the ROC curve was also plotted. The Xgboost algorithm performed well in all of these evaluation metrics, demonstrating that the proposed smote-Xgboost based framework has a significant advantage in predicting heart disease. In addition, we estimated the feature importance and related scoring of the four algorithms to provide ideas for further optimization of the algorithms.

5. Conclusions

In this research, we present a smote-Xgboost-based methodology for heart disease prediction. Firstly, an approach for choosing features using information gain is proposed, and then the hybrid Smote-Enn algorithm is used to process unbalanced datasets. Finally, the processed HDD dataset is used for model training. In the experimental evaluation, we compare the Xgboost algorithm with five baseline algorithms. The outcomes demonstrate that the model suggested in this research performs exceptionally well across all four evaluation indicators, with prediction accuracy of 93.44%. In addition, we also count the feature importance of the selected algorithm, which has important implications in terms of heart disease prediction.

In future work, we will mix multiple effective machine learning techniques and combine cutting-edge data processing techniques to build real-time and reliable heart disease diagnosis models.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y. and J.G.; software, J.G.; formal analysis, J.Y. and J.G.; investigation, J.Y.; writing—original draft preparation, J.Y. and J.G.; writing—review and editing, J.Y. and J.G.; visualization, J.G.; supervision, J.Y.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This article is supported in part by the Humanities and Social Science Fund of Ministry of Education of China (Grant No. 21YJCZH197); the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (Grant No. 2020L0252) and Shanxi Undergraduate Training Program for Innovation and Entrepreneurship (Grant No. 20220338).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments.

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

References

1. Cardiovascular Diseases. Available online: <https://www.who.int/health-topics/cardiovascular-diseases/> (accessed on 10 September 2022).
2. Shah, S.; Shah, F.; Hussain, S.; Batool, S. Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods. *Comput. Electr. Eng.* **2020**, *84*, 106628. [\[CrossRef\]](#)
3. Che, C.; Zhang, P.; Zhu, M.; Qu, Y.; Jin, B. Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 184. [\[CrossRef\]](#)
4. Hoodbhoy, Z.; Jiwani, U.; Sattar, S.; Salam, R.; Hasan, B.; Das, J. Diagnostic Accuracy of Machine Learning Models to Identify Congenital Heart Disease: A Meta-Analysis. *Front. Artif. Intell.* **2021**, *4*, 197. [\[CrossRef\]](#)
5. Wang, Z.; Chen, L.; Zhang, J.; Yin, Y.; Li, D. Multi-view ensemble learning with empirical kernel for heart failure mortality prediction. *Int. J. Numer. Methods Biomed. Eng.* **2020**, *36*, e3273. [\[CrossRef\]](#)
6. Modepalli, K.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20–22 January 2021.
7. Joo, G.; Song, Y.; Im, H.; Park, J. Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). *IEEE Access* **2020**, *8*, 157643–157653. [\[CrossRef\]](#)
8. Li, J.; Haq, A.; Din, S.; Khan, J.; Khan, A.; Saboor, A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access* **2020**, *8*, 107562–107582. [\[CrossRef\]](#)
9. Ali, F.; El-Sappagh, S.; Islam, S.M.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf. Fusion* **2020**, *63*, 208–222. [\[CrossRef\]](#)
10. Rahim, A.; Rasheed, Y.; Azam, F.; Anwar, M.; Rahim, M.; Muzaffar, A. An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. *IEEE Access* **2021**, *9*, 106575–106588. [\[CrossRef\]](#)
11. Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access* **2021**, *9*, 39707–39716. [\[CrossRef\]](#)
12. Khurana, P.; Sharma, S.; Goyal, A. Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques. In Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021, Matsue, Japan, 18–22 October 2021.
13. Ashri, S.E.A.; El-Gayar, M.M.; El-Daydamony, E.M. HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. *IEEE Access* **2021**, *9*, 146797–146809. [\[CrossRef\]](#)
14. Bashir, S.; Almazroi, A.; Ashfaq, S.; Almazroi, A.; Khan, F. A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction. *IEEE Access* **2021**, *9*, 130805–130822. [\[CrossRef\]](#)
15. Odhiambo Omuya, E.; Onyango Okeyo, G.; Waema Kimwele, M. Feature Selection for Classification using Principal Component Analysis and Information Gain. *J. Biomed. Inform.* **2021**, *174*, 114765. [\[CrossRef\]](#)
16. Le, T.; Lee, M.; Park, J.; Baik, S. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry* **2018**, *10*, 79. [\[CrossRef\]](#)
17. Vandewiele, G.; Dehaene, I.; Kovács, G.; Sterckx, L.; Janssens, O.; Ongenaes, F.; Backere, F.D.; Turck, F.D.; Roelens, K.; Decruyenaere, J.; et al. Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artif. Intell. Med.* **2021**, *111*, 101987. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *J. Biomed. Inform.* **2020**, *107*, 103465. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Budholiya, K.; Shrivastava, S.; Sharma, V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *34*, 4514–4523. [\[CrossRef\]](#)
20. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
21. Asadi, S.; Roshan, S.; Kattan, M.W. Random forest swarm optimization-based for heart diseases diagnosis. *J. Biomed. Inform.* **2021**, *115*, 103690. [\[CrossRef\]](#)
22. Bansal, M.; Goyal, A.; Choudhary, A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decis. Anal. J.* **2022**, *3*, 100071. [\[CrossRef\]](#)
23. Książek, W.; Gandor, M.; Pławiak, P. Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Comput. Biol. Med.* **2021**, *134*, 104431. [\[CrossRef\]](#)
24. Ghiasi, M.M.; Zendehboudi, S.; Mohsenipour, A. Decision tree-based diagnosis of coronary artery disease: CART model. *Comput. Methods Prog. Biomed.* **2020**, *192*, 105400. [\[CrossRef\]](#)
25. Chen, S.; Webb, G.I.; Liu, L.; Ma, X. A novel selective naïve Bayes algorithm. *Knowl.-Based Syst.* **2020**, *192*, 105361. [\[CrossRef\]](#)