ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.167

An Ensemble Learning Model for Heart Disease Prediction

Qingli Li[1], Yagang Wang[1], Shizhang Hu[1], Xiangchun Jin[1]

[1]School of Computer Science and Technology, Xi'an University of Posts & Telecommunications, Xi'an, China

lql1112575@163.com, wangyg@xupt.edu.cn, 18879405097@163.com, hyangcun@163.com

Abstract. Heart disease is a common disease that seriously threatens the health of mankind, especially the middle-aged population and the elderly. It is characterized by high prevalence, disability rate, and mortality rate. The patients are never completely cured even with the most advanced treatment. However, the prediction of the incidence in advance and the enabling of doctors on formulating scientific treatment plans would significantly improve the cure rate. In light of the current situation, this paper proposes a cardiovascular disease prediction model based on ensemble learning, which integrates several classical machine learning algorithms such as AdaBoost, Random Forest(RF), Support Vector Machine(SVM), Lightgbm, and Gradient Boosting Decision Tree(GBDT). The first four algorithms with low correlation are constructed as base learners and then ensemble into the meta learner Gradient Boosting Decision Tree to build an ensemble learning model. The excellence of the ensemble model is evaluated from the evaluation indexes such as accuracy, precision, and recall. The experimental results show that the accuracy rate of the ensemble model is 90%, the precision rate is 90.1%, and the recall rate is 90.66%. Compared with a single machine learning prediction model, the maximum improvement is 4%. The model will effectively assist doctors in making more accurate predictions of patients' physical conditions and carrying out scientific treatment.

Keywords: heart disease, Adaboost, Random Forest, Support Vector Machine, Lightgbm, GBDT, Ensemble learning

1. Introduction

Heart disease is the leading cause of death in the world. It is estimated that 17.9 million people die every year, accounting for 31% of the total deaths in the world [1]. With the development of the economy and society, more and more factors are causing cardiovascular diseases, including unreasonable eating habits and irregular work and rest. As the most populous country in the world, China has a gradual increase in the incidence rate of heart diseases and is showing a younger trend [2,3]. At the same time, due to the unbalanced development of China at this stage, the medical resources in remote areas are insufficient [4]. Under this background, using machine learning methods to assist doctors in the scientific and accurate treatment of patients can improve treatment efficiency and reduce the waste of public medical resources.

People with heart disease or high risk (due to more risk factors, such as hypertension, diabetes, hyperlipidemia, or established conditions) need early detection and management. Machine learning models can be of great help. As early as 1976, Wraith et al. of Stanford University developed the MYCIN system to help doctors diagnose patients [5]. However, due to the limited computing performance of early computers, the accuracy of the MYCIN system is limited, and the function is relatively scarce. The heart failure survival score (HFSS) model was proposed in 1997. It is the first multivariable-based survival model for patients with heart failure and finds the most important influencing factors affecting the death of patients with heart failure [6]. However, due to its long history, the performance of accuracy is limited. In recent years, due to the rapid development of computers and the rise of machine learning, machine learning has been widely used in medicine and bioengineering. More and more researchers mine patients' medical record information through machine learning methods, find disease-related information from medical record data, and use this information to establish disease prediction models [7]. Lizhi Miao et al used support vector machines and random forests in the prediction and analysis of breast cancer risk[8]. Fengshui Miao et al. used random forests and XG-Boost to predict the risk of diabetes [9]. Yingxu Zhao et al. used Lightgbm to predict the hospitalization expenses of senile dementia [10]; Amin et al. completed the prediction of heart disease using

Support Vector Machine [11]. Mark Stampehl et al.conducted a risk assessment on mortality of patients with heart failure. They used a classified regression tree, complete logistic regression, and stepwise logistic regression methods in the model [12]. Among them, stepwise logistic regression performed best, and the ROC of the model reached 0.74-0.75. Sho Suzuki and others used multiple logistic regression analysis with stepwise variable selection to select mortality prediction variables and determine the weighted score to find the most important influencing factors [13].

In a word, machine learning algorithm has made significant progress in medicine. Constructing relevant models can assist doctors' diagnosis, and better help doctors understand the situation of patients and carry out the scientific treatment.

Based on the existing research, this paper proposes a heart disease prediction model based on ensemble learning. Firstly, the data is normalized to the maximum and minimum, and then four base learners with low correlation, such as AdaBoost, random forest, support vector machine, and Lightgbm, are constructed. Finally, the four base learners are fused and ensembled into the meta learner GBDT to create the blended learning model. Compared with a single learner, the model has higher accuracy, precision, and recall and can more effectively assist doctors in diagnosis and treatment.

2. Data

2.1. Data Introduction

The data used in this research can be obtained from Kaggle competitions. This dataset was created by combining different datasets already available independently but not combined before. In this dataset, five heart datasets are connected over 11 standard features, making it the largest heart disease dataset available for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

There are 1190 observations in this dataset, of which 272 observations are repeated, so the final dataset has 918 observations. Each sample data has 12 attributes, of which the 12th attribute represents whether it is sick or not. The feature attributes in the final dataset are shown in table 1:

Table 1 Dataset Feature Attributes

Feature Name	Feature Description	Feature Type
Age	age of the patient	Number
Sex	sex of the patient(M, F)	Classification
ChestPainType	chest pain type (TA,ATA,NAP,ASY)	Classification
RestingBP	resting blood pressure	Number
Cholesterol	serum cholesterol	Number
FastingBS	fasting blood sugar[1:FastingBS > 120 mg/dl, 0: otherwise]	Classification
RestingECG	resting electrocardiogram results(Normal,ST,LVH)	Classification
MaxHR	maximum heart rate achieved	Number
ExerciseAngina	exercise-induced angina(Y, N)	Classification
Oldpeak	oldpeak = ST	Number
ST_Slope	the slope of the peak exercise ST segment(Up, Flat, Down)	Classification
Heart Disease	output class [1: heart disease, 0: Normal]	Classification

2.2. Data preprocessing

One of the main purposes of data preprocessing is to filter the characteristic data needed for research from the data set. In the process of data preprocessing, it is found that the data in the data set is in good condition, and there are no common problems such as missing values and outliers. At the suggestion of professional doctors, the characteristic field of chestpainttype in the data set was removed in this study. The reason is that the type classification of chest pain mainly depends on the patient's self-description of its pain location and nature, and there are many inducing and interfering factors of chest pain, so it should not be used as reference data. It can be seen from Table 1 that the feature types in the data sets used in this study are not the same, including binary and multivalued classes.

Another main purpose of data preprocessing is to convert the data in the data set into data recognizable by the algorithm model [14]. As can be seen from table 1, the feature types in the dataset used in this study are not the same, including binary and multivalued classes and numeric and character types. For example, the sex field is binary data, and the data type is a character type, while the age field is multivalued data, and the data type is a number type. In the process of data preprocessing, the male in the gender field is set to 0 and the female is set to 1. Fields of other binary classes, and so on. Set Normal to 0, ST to 1, and LVH to 2 in the RestingECG field Fields of other multivalued classes, and so on.

After all the data in the data set are converted into digital types, the data of each attribute is normalized. Data normalization is to reduce the data to a specific threshold according to a certain proportion. The reason is that there are different dimensions for different data, and the calculation is more complex. In this study, the max-min normalization method is used to convert all the data into numbers between [0,1]. The calculation formula is shown in formula (1):

$$x = \frac{x - V_{min}}{V_{max} - V_{min}} \tag{1}$$

2.3. Data feature correlation analysis

Data set features need to be analyzed. Using the relationship between features can better understand the relationship between data set features, called correlation in statistics[15]. Using the thermal diagram to visualize the relationship between features, we can see the correlation between various features. Each color block represents the correlation degree of the features of abscissa and ordinate. The darker the color, the greater the correlation coefficient, the stronger the correlation, the deeper the degree of redundancy, and the lighter the color; it is just the opposite. Figure1 shows the thermodynamic diagram of correlation analysis between the first 10 features in this data set. It can be seen from the figure that the correlation between each feature is weak, and this means that the dataset does not need to eliminate any features.

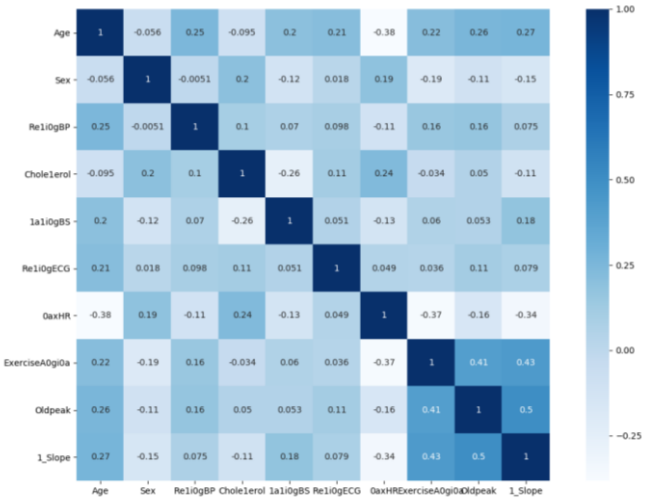

Fig.1 Characteristic correlation heat map

3. Methods

3.1. Ensemble learning

biased; they perform well in some aspects and poorly in others. To overcome this problem, Ensemble learning came into being. Ensemble learning obtains a better and more comprehensive robust supervision model by combining multiple weak supervision models. The main idea is that most weak classifiers correct the wrong prediction of a weak classifier and improve the prediction accuracy. There are three primary forms of ensemble learning: bagging, boosting, and stacking.

Ensemble learning performs well on data sets of different sizes. When the data set is large, it can be divided into multiple small data sets to learn multiple models for combination. When the data set is small, multiple small data sets are obtained by sampling, and then various models are trained for variety.

3.2. Stacking Model fusion

Model fusion combines some relatively simple algorithms to retain the advantage of the low variance of these algorithms. On this basis, it can introduce complex models to expand the prediction space of simple algorithms. Ensemble learning model fusion can reduce standard deviation and deviation. This paper mainly uses the stacking method for model fusion. Stacking ensemble learning is a strategy of heterogeneous integration. Heterogeneous integration integrates several different types of base classifiers into a robust classifier to improve the generalization ability of solid classifiers. A two-layer model can generally summarize the stacking fusion model. The first layer is composed of multiple base learners, and the second layer is meta learners. The stacking model is obtained by fusing numerous base learners in the first layer and meta learners in the second layer. The stacking method trains primary learners from the original data set and then uses the new data set to teach secondary learners.

By understanding the stacking method, the integration effect depends on two aspects: one is the prediction effect of the base classifier. Generally, the better the prediction effect of the base classifier is, the better the prediction effect of the ensemble learning model is; other is that there must be some differences between base classifiers because the main concerns of each model are different so that each base learner can give full play to its advantages.

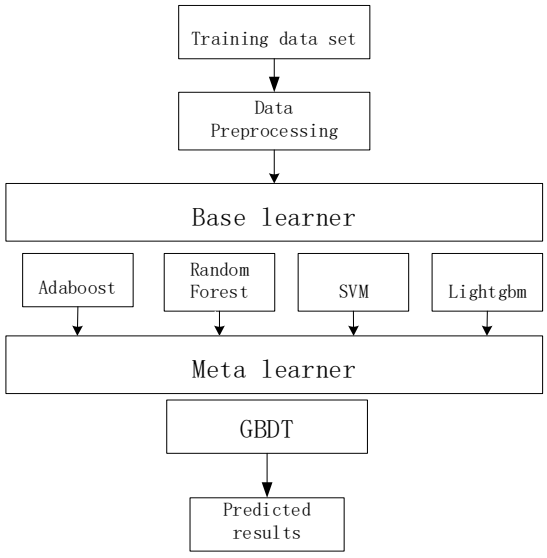

Fig.2 Ensemble learning model based on stacking

In this paper, the classification is based on the idea of stacking model fusion process, and the model process is shown in Figure 2:

1) Firstly, the training set is divided into a training set and test set according to the ratio of 7:3.

2) The split training set is sent to four different base learners AdaBoost, Random Forest, SVM, and Lightgbm in the first layer for training.

3) The training results of the first layer are spliced and transmitted to the meta learner: GBDT, which is trained again, and the final training results are obtained.

4) After completing the above three steps, the test set data to be predicted is sent to the model for prediction.

The Ensemble learning model, base learner, and meta learner used in this study are implemented by Python code, and the Sk-learn library is used at the same time. Sk-learn is a tool often used in machine learning research. It includes supervised learning and unsupervised learning methods. Using Sk-learn can greatly save the time and workload of writing code so that researchers can focus more on the choice of methods.

3.3. Base learner

Based on the 10 characteristic attributes in the data, an ensemble learning model is constructed to classify and predict heart disease. The primary learner in the ensemble model adopts the following schemes:

1) AdaBoost is an iterative algorithm. Its core idea is to train different classifiers (weak classifiers) for the same training set and then collect them to form a more robust final classifier (strong classifier).

2) Random forest is a classifier containing multiple decision trees. It uses various decision trees to train and predict samples. Most of the time, it is trained with the "bagging" method. It uses randomly put back selected training data, then constructs a classifier, and finally combines the learned models to increase the overall effect.

3) Support Vector Machine (SVM) is a typical binary classification model. Its basic model is the linear classifier with the most considerable interval defined in the feature space. The basic idea is to solve the separation hyperplane, which correctly divides the training data set and has the most significant geometric interval.

4) Gradient lifting decision tree(Lightgbm) is a decision tree algorithm based on the histogram. The basic idea is to improve the gradient of the decision tree when it is operated.

3.4. Meta learner

The prediction deviation is again transferred to the meta learner: gradient descent tree (GBDT) for Ensemble model fusion learning according to the base learner trained above.

GBDT (gradient boosting decision tree), also known as MART (multiple additive regression tree), is an iterative decision tree algorithm. The algorithm is composed of numerous decision trees, and the conclusions of all trees are accumulated to make the final answer. It is an algorithm with strong generalization ability.

4. Experiment

4.1. Experimental Environment

Experimental hardware environment: the CPU is Intel Core i5-10400, the GPU is NVIDIA RTX 2060 gdr6 6GB, the memory is DDR4 32g, and the operating system is window 10 64 bit. The experimental software environment is Python version 3.7 and pycharm professional 2021.2 integrated development environment.

4.2. Metrics

When using machine learning algorithms for regression, classification, or clustering tasks, there are many quantitative indicators to test the effect of the machine learning model, which are called evaluation indicators.

Actual / Predict	0	1
0	TN	FN
1	FP	TP

Fig.3 Confusion Matrix

The table summarizing the prediction results of the classification model is usually called the confusion matrix. Taking the second classification as an example, the confusion matrix is shown in Figure 3.

Where TP = true positive; FP = false positive; FN = false negative; TN = true negative. TP represents the number of samples that are positive and predicted to be positive, FP represents the number of samples that are negative but predicted to be positive, FN represents the number of samples that are positive but predicted to be negative. TN represents the number of negative and predicted to be negative samples.

(1) Accuracy: Accuracy is the proportion of correctly classified samples to the total number of pieces, as shown in formula 2:

$$Accuracy = \frac{n_{correct}}{n_{total}} \tag{2}$$

The number of correctly classified samples is incorrect, and the total is the total number of pieces. Combined with the confusion matrix, the formula can also be expressed as shown in 3:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Accuracy is the most straightforward and intuitive evaluation index in classification problems, but it has obvious defects. For example, if 99% of the samples are positive samples, the classifier only needs to predict positively all the time to get 99% accuracy, but its actual performance is deficient. In other words, when the proportion of samples in different categories is very uneven, the types with a large proportion often become the main factor affecting the accuracy. Therefore, we should take the accuracy rate as the only evaluation index and pay attention to other evaluation indexes at the same time.

(2) Precision: Precision refers to the proportion of samples that are positive in the instances predicted to be positive by the model. as shown in formula 4:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

The precision is the TP / (FP + FP) ratio, where TP is the number of authenticity and FP is the number of false positivity. The precision rate is intuitively the ability of the classifier not to mark negative samples as positive samples.

(3)Recall: Recall refers to the proportion of the predicted positive samples in the actual positive examples, as shown in formula 5:

$$Reacll = \frac{TP}{TP + FN} \tag{5}$$

The recall is the ratio TP / (TP + FN), where TP is the number of faithful and FN is the number of false negativity. Recall rate is intuitively the ability of the classifier to find all positive samples. The best value of the recall rate is one, and the worst value is 0.

(4)ROC curve: ROC curve is the abbreviation of the receiver operating characteristic curve. The abscissa of the ROC curve is the false positive rate (FPR); The ordinate is the true positive rate (TPR). The calculation methods of FPR and TPR are 6-5 and 6-6, respectively:

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

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

The abscissa of each point on the ROC curve is FPR, and the ordinate is TPR.

(5)AUC: AUC (area under the curve) is the area under the ROC curve, which can quantitatively reflect the model performance measured based on the ROC curve. The value of AUC is generally between 0.5 and 1.

The larger the AUC, the more likely the classifier is to rank the positive samples before the actual negative examples, that is, make a correct prediction.

4.3. Experimental Results

The final data set used in this study contains 918 samples, each with 10 characteristics to predict whether there is heart disease. To verify whether the ensemble learning model has a better prediction effect on the data set, this study experiments the ensemble learning model with the basic learner and meta learner respectively. The evaluation indexes are accuracy, precision, recall, and AUC, and compares the final results. After many experiments, the final experimental results are shown in Table 2.

Table.2 Comparison of results of different algorithms

Model	Accuracy	Precision	Recall	AUC
AdaBoost	86.69%	86.46%	85.59%	0.90
Random Forest	88.26%	88.32%	87.90%	0.91
SVM	89.08%	89.15%	88.65%	0.92
Lightgbm	86.39%	86.43%	85.02%	0.90
GBDT	86.08%	86.05%	85.75%	0.90
Stocking	90.00%	90.10%	90.66%	0.94

As shown from Table 2, compared with AdaBoost, random forest, SVM, Lightgbm, and GBDT, the evaluation indexes such as accuracy, precision, and recall of the ensemble model proposed in this study are the largest. It reached 90%, 90.1% and 90.66% respectively. The AUC value was also the maximum value, reaching 0.94. This shows that the ensemble learning model has improved in varying degrees in all aspects of performance. To better display the AUC value, ROC curves of different algorithms are compared, as shown in Figure 4.

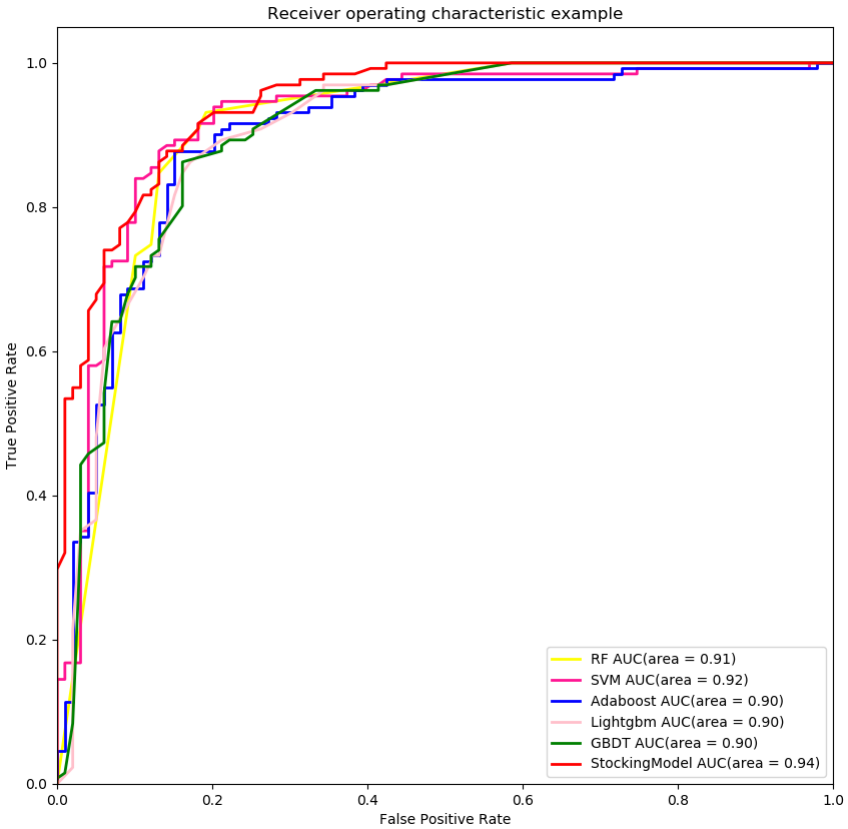

Fig.4 ROC curves of different algorithms

5. Conclusions

The experiments in this study are conducted on the most critical heart disease data set provided by the Kaggle competition for research purposes. While previous studies mostly use a single machine learning algorithm with insufficient accuracy, a heart disease prediction model based on ensemble learning is proposed, which incorporates AdaBoost, Random Forest, Support Vector Machine Lightgbm, GBDT, and other classical machine learning algorithms. Positive results are obtained by preprocessing the data set, analyzing the characteristics, and using the ensemble learning model to predict. Compared with a single machine learning algorithm, it significantly improves accuracy, precision, recall, and AUC value. The model will effectively assist doctors in making more accurate predictions of patients' physical conditions and carrying out scientific treatment.

6. References

[1] Yancy C W, Jessup M, Bozkurt B, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American college of cardiology foundation/American heart association task force on practice guidelines[J]. Journal of the American College of Cardiology, 2013, 62(16): 147-239.

[2] Ding W J.Research on classification algorithm in heart disease prediagnosis[D].Xi'an: Xidian University, 2019.

[3] Li X Q. Research on heart disease prediction method based on convolutional neuralnetwork[D]. Harbin: Northeast Forestry University, 2019.

[4] Kuang X Y, Hou H R. Current status of risk prediction models based on electronic medical record data in clinical care[J]. Journal of Nursing,2020,27(16):21-24.

[5] Wraith S M, Aikins J S, Buchanan BG, et al. Computerized consultation system for selection of antimicrobial therapy[J].American Journal of Hospital Pharmacy,1976,33(12):1304-1308.

[6] Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC guidelines for diagnosing and treating acute and chronic heart failure[J]. European Journal of Heart Failure, 2016, 18(8): 891-975.

[7] Zhao M D, Sun J A.Research progress of machine learning in cardiovascular disease diagnosis[J].Beijing Biomedical Engineering, 2020,39(2): 208-214.

[8] Miao L Z, Diao J Y, Lou C, et al. Analysis of breast cancer risk prediction based on Spark and Random forest[J]. Computer Technology and Development, 2019, 29(8):142-146.

[9] Miao FS, Li Y, Gao C, et al.Diabetes prediction method based on CatBoost algorithm[J]. Computer System Applications, 2019,28(9):215-218.

[10] Zhao Y X, Bao Z Q, Gao S, et al.Prediction and comparison of medical insurance hospitalization costs considering Alzheimer's disease based on machine learning model[J]. Insurance Research, 2020,9(5):64-76.

[11] Amin U H, Li J P, Muhammad H M.A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms[J]. Mobile Information Systems,2018(12):1-21.

[12] Stampehl M, Friedman HS, Navaratnam P, et al. Risk assessment of post-discharge mortality among recently hospitalized medicare heart failure patients with reduced or preserved ejection fraction[J]. Current Medical Research and Opinion, 2020, 36(2): 179-188.

[13] Suzuki S, Motoki H, Kanzaki Y, et al. A predictive model for 6-month mortality in elderly patients with heart failure[J]. International Heart Journal, 2020, 61(2): 325-331.

[14] Nikhil G, Priya C, Shravani P. Heart disease prediction using classification techniques: a comparative study[J]. International Journal of Management, IT and Engineering,2019,9(7):77-89.

[15] Aguero A G, Sayago A, Gonzalez A G. The correlation coefficient: An overview[J]. Critical Reviews in Analytical Chemistry,2006,36(1):41-59.