A Stacking-Based Heart Disease Classification Prediction Model

Pan Liu[1], Li Li[2,*], and Yihao Li[3]
[1]Faculty of Business Information, Shanghai Business School, Shanghai 201400, China
[2]School of Literature and Journalism, Yichun University, Yichun, China
[3]School of Information and Electrical Engineering, Ludong University, Yantai, China
[*]corresponding author: 5571990@qq.com

Abstract—Heart disease is one of the leading causes of death globally, and early diagnosis and prevention of heart disease are of great significance. The paper provides a correlation analysis of the quantitative and qualitative variables on a heart disease dataset. Based on the Pearson correlation coefficient matrix, the paper determines the base classifiers and uses logistic regression as the meta-classifier to construct a stacking ensemble learning model for heart disease classification prediction. To address the issue of data imbalance, cost-sensitive learning is further introduced, and a stacking heart disease classification prediction model based on threshold optimization is established. The experimental result shows that the stacking heart disease classification prediction model achieves a classification accuracy of 90.16% and demonstrates good generalization ability. Our work provides guidance for further research on heart disease classification prediction.

Keywords-stacking integrated learning; Pearson correlation coefficient matrix; cost-sensitive learning; threshold optimization; prediction of heart disease

I. INTRODUCTION

Heart disease is one of the leading causes of death globally. According to the report on cardiovascular health and disease report in China 2021 [1], Chinese rural and urban areas accounted for 46.74% and 44.26% of cardiovascular disease (CVD) deaths, respectively, in 2019. The current number of individuals affected by cardiovascular disease is approximately 330 million. The prevalence and mortality rates of cardiovascular disease in China are still on the rise, with heart disease patients constituting the majority. However, the traditional diagnostic methods [2] for heart disease heavily rely on the doctor's experience and expertise, which can lead to significant variations in diagnosis among different physicians. Therefore, machine learning techniques [3] have been proposed to assist in the diagnosis of heart disease in the practice.

In recent years, various scholars have conducted extensive research on the diagnosis of heart disease. Wisaeng [4] predicted the diagnosis of heart disease using feature selection and the k-nearest neighbors algorithm, demonstrating promising predictive performance. Yadav and Pal [5] proposed a classification prediction method based on chi-square feature selection and random forest, which can assist healthcare professionals in predicting heart diseases. Ramachandran et al. [6] utilized the Framingham model to predict the incidence of heart disease, providing a scientific basis for heart disease prevention. Katarya and Srinivas [7] reviewed the research on SVM for heart disease prediction. Khan and Kim [8] employed LSTM for classifying and predicting heart disease. Wang [9] presented a heart disease prediction model based on XGBoost and K-means, verifying its feasibility and effectiveness through empirical studies. Guo et al. [10] proposed an ensemble learning model based on the Stacking algorithm, which can effectively address clinical diagnostic issues related to heart disease. Sharma [11] designed and developed an ensemble classifier for heart disease that integrates Support Vector Machine (SVM), K–Nearest Neighbor (K-NN), and Weighted K-NN.

In summary, scholars have employed various machine learning and data mining methods for the diagnosis and prediction of heart disease. Their aim is to enhance the accuracy of the models and achieve better diagnostic results, providing precise and reliable assistance for heart disease diagnosis. However, most of the existing studies have overlooked the issue of cost sensitivity, which is more prevalent in the medical field. Therefore, the paper introduces the concept of cost-sensitive learning to further investigate the classification prediction of heart disease. Then, the paper proposes a stacking-based heart disease classification prediction model that incorporates cost-sensitive learning, aiming to improve the accuracy and generalization capability of the model. Our research provides a direction for further studies in heart disease classification prediction.

II. DATA ANALYSIS

We select a publicly available dataset from PaddlePaddle[1] for a specific research study. The dataset consists of data from 303 patients who underwent heart disease examinations, including 13 influential features and 1 target feature. The specific feature list and their descriptions are shown in Table 1.

The dataset consists of a total of 303 samples, with 138 samples labeled as positive (target = 0) and 165 samples labeled as negative. The class distribution is imbalanced, with fewer positive samples. In Table 1, among the features, "sex", "cp", "fbs", "restecg", "exang", "slope", "thal", and "target" are categorical variables, while "age", "trestbps", "chol", "thalach", "oldpeak", and "ca" are quantitative variables.

[1] https://aistudio.baidu.com/aistudio/datasetdetail/108029

2767-6684/23/$31.00 ©2023 IEEE
DOI 10.1109/DSA59317.2023.00076

TABLE 1. LIST OF DATA SET FEATURE

Feature	Meaning	Description
age	Age	Age of the patient
sex	Sex	0 = Male; 1 = Female
cp	Chest Pain Type	0 = Typical angina; 1 = Atypical angina; 2 = Non-anginal pain; 3 = Asymptomatic
trestbps	Resting Blood Pressure	Blood pressure at rest measured in mm Hg
chol	Cholesterol	Serum cholesterol level in mg/dl
fbs	Fasting Blood Sugar	0 = Fasting blood sugar < 120 mg/dl; 1 = Fasting blood sugar > 120 mg/dl
restecg	Resting ECG Results	0 = Normal; 1 = Abnormal ST-T wave; 2 = Possible or definite left ventricular hypertrophy
thalach	Maximum Heart Rate	Maximum heart rate achieved
exang	Exercise-Induced Angina	0 = No; 1 = Yes
oldpeak	ST Depression	ST depression induced by exercise relative to rest
slope	Slope of ST Segment	0 = Upsloping; 1 = Flat; 2 = Downsloping
ca	Number of Major Vessels	Number of major vessels colored by fluoroscopy
thal	Thalassemia	0 = Uncertain; 1 = Normal; 2 = Fixed defect; 3 = Reversible defect
target	Disease Status	0 = Healthy; 1 = Diseased

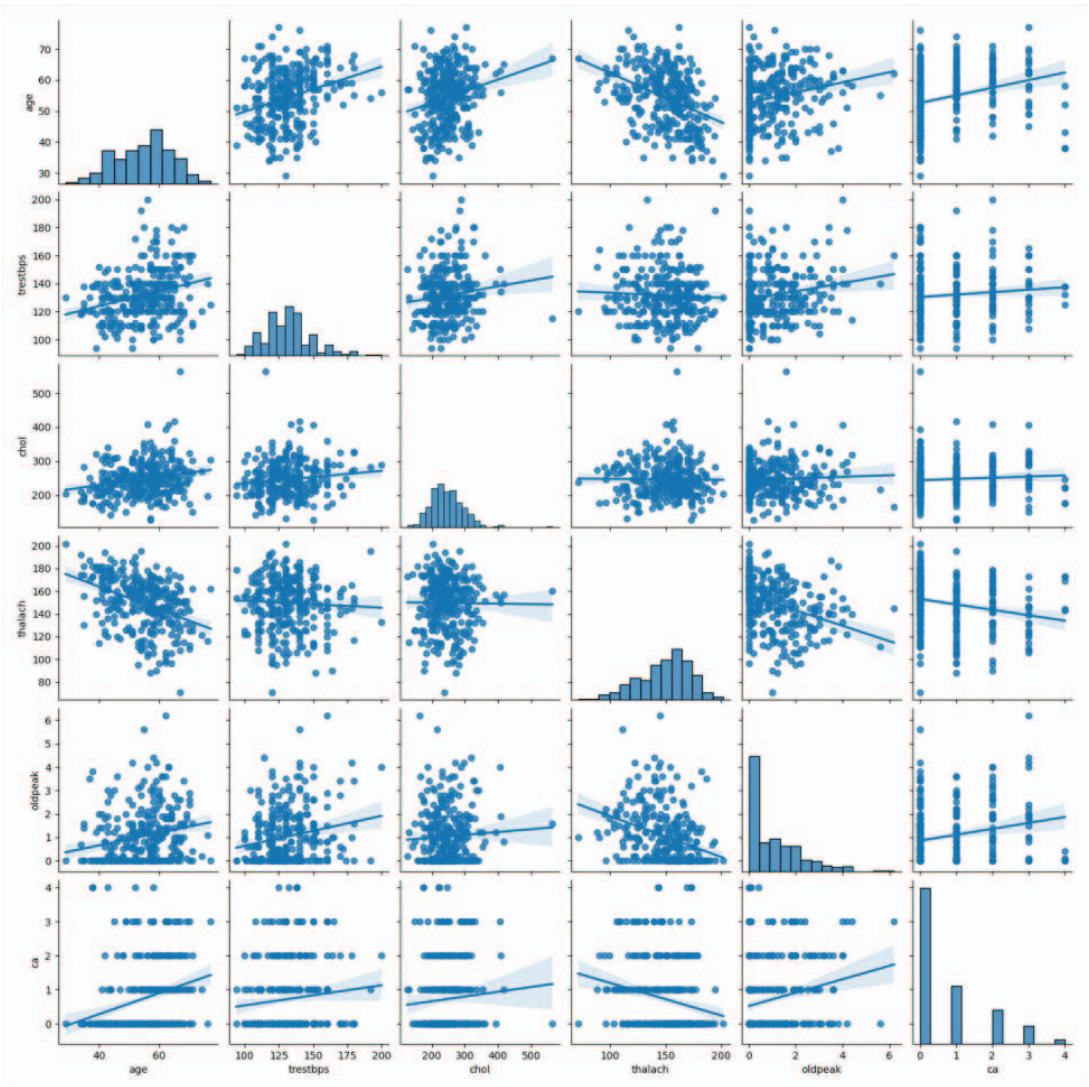

Figure 1. Diagram of relationship of quantitative variables

A. Variable Correlation Analysis

We create scatter plots between all the quantitative variables in Table 1 to examine the relationships between them. The scatter plots provide an initial assessment of whether there are connections between each pair of variables, as shown in Figure 1. The straight lines represent the best-fit linear regression lines for each pair of variables. By observing these fitted lines, it can be concluded that the linear relationships between the variables are weak, and there is no clear evidence of significant correlation between them.

Then, we compute the Pearson correlation coefficients for each quantitative relationship and generate a heatmap to visualize the results, as shown in Figure 2. It can be observed that all the correlation coefficients fall within the range of (-0.4, 0.4). This indicates that there is no strong correlation between the quantitative variables. Therefore, there is no need to remove redundant variables as there is no significant correlation among them.

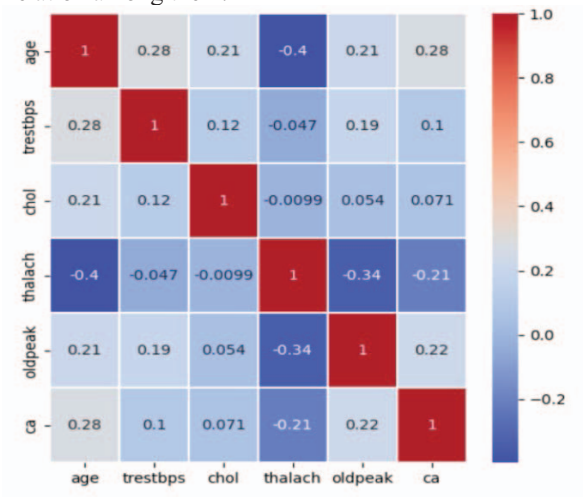

Figure 2. A heatmap of correlation coefficients in Table 1

B. Analysis of qualitative variables

Figure 3 shows pie charts for all the categorical variables in the dataset. It can be observed that in the "target" variable, the proportion of diseased individuals is 54.5%, while the proportion of healthy individuals is 45.5%. This indicates the presence of data imbalance, where the number of diseased individuals is higher than the number of healthy individuals.

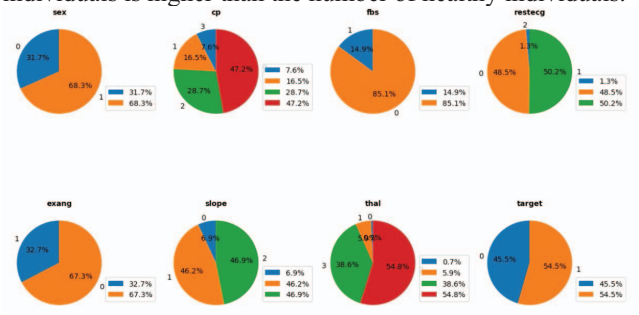

Figure 3. Pie charts of qualitative variables

To further analyze the factors influencing heart disease, some pie charts specifically for the diseased individuals are plotted in Figure 4. Based on the observations, the following preliminary conclusions can be drawn:

1) The proportion of females with heart disease is higher than that of males.
2) Non-anginal chest pain type has a higher proportion of individuals with heart disease.
3) Fasting blood sugar level is an important indicator for identifying heart disease.
4) The presence of ST-T wave abnormalities in the electrocardiogram is associated with a higher proportion of individuals with heart disease.
5) Exercise-induced angina has a lower proportion of individuals with heart disease. If angina occurs during exercise, other diseases should be considered as a priority.
6) A downward sloping ST segment during exercise is associated with a higher proportion of individuals with heart disease.
7) Reversible defects have a higher proportion of individuals with heart disease.

These conclusions provide initial insights into the factors that may contribute to heart disease based on the observed proportions in the pie chart.

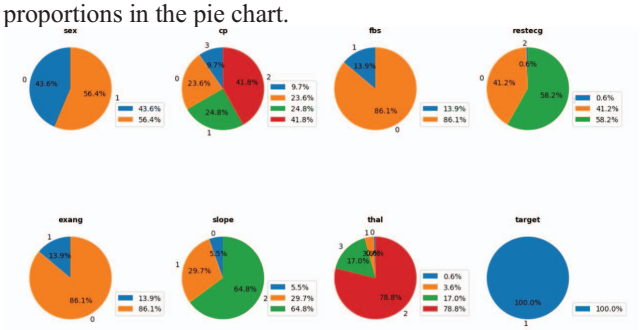

Figure 4. Pie charts of disease data

C. Data normalization

Due to the presence of different scales and magnitudes in the dataset, it is necessary to perform data normalization before building the model. In this study, we adopt the method of Min-Max normalization, which maps the range of the original data to [0, 1]. This approach effectively eliminates the scale differences among different variables. The specific formula is as follows:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

III. STACKING-BASED HEART DISEASE CLASSIFICATION PREDICTION MODEL

A. Stacking ensemble learning model framework

Stacking, proposed by Wolpert [12] in 1992, consists of two parts: base learners and a meta-learner. It is essentially a serial multi-level learning system, typically consisting of two layers. The first layer is the base learners, composed of several different machine learning algorithms. The second layer is the meta-learner, which takes the output from the first layer as input and produces the final prediction.

To generate training samples for the meta-learner and prevent overfitting in classic stacking ensemble methods [13], k-fold cross-validation is commonly used. K-fold cross-validation randomly divides the dataset into k mutually exclusive and equally sized subsets. The model is trained on k-1 subsets and tested on the remaining subset. This process is repeated for all possible k selections, and the results are used as new sample features.

The specific steps of the stacking ensemble learning model in the paper are as follows:

1) Input the dataset $D = \{(y_i, x_i); i = 1, 2, 3, \dots, 5m)\}$ and split it into the original training set $D_{train} = \{(y_i, x_i); i = 1, 2, 3, \dots, 4m)\}$ and the original test set $D_{test} = \{(y_i, x_i); i = 1, 2, 3, \dots, m)\}$ in a 4:1 ratio.
2) Apply k-fold cross-validation on the original training set D_{train} to train n base classifiers $C_1, C_2, \dots, C_n$, and generate the secondary training set $D'_{train} = \{(y_i, z_{1i}, z_{2i}, \dots, z_{ni}); i = 1, 2, 3, \dots, 4m\}$, where y_i is the label of the corresponding data in the original training set, and $(z_{ji}, z_{ji}, \dots, z_{ji}); i = 1,2,3, \dots, 4m$ are the classification results of the j-th base classifier on D_{train}. Stacking typically uses the predicted class labels, but this model uses predicted class probability values, which are more suitable for further research on the problem.
3) During the k-fold cross-validation training process of the base classifiers, the original test set is also evaluated k times, and the average is taken to generate the secondary test set $D'_{test} = \{(y_i, z_{1i}, z_{2i}, \dots, z_{ni}); i = 1,2,3, \dots, m\}$.
4) Train the meta-classifier model on the secondary training set D'_{train} and test it on the secondary test set D'_{test}.

The model framework is shown in Figure 5.

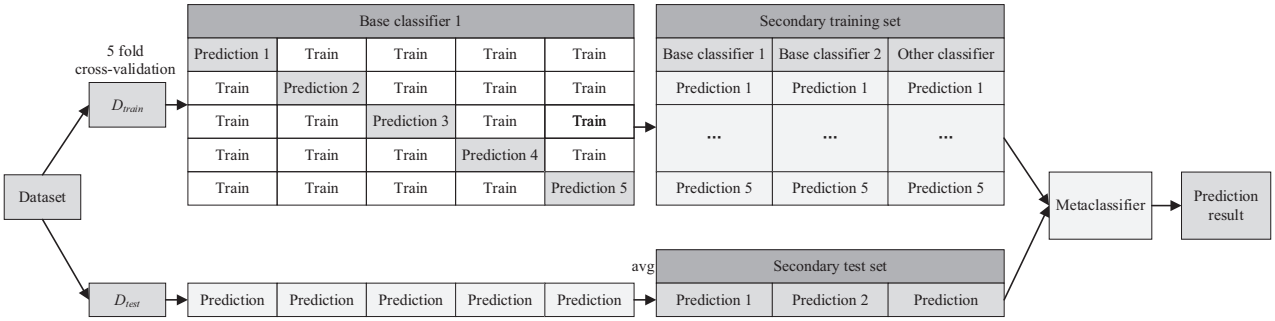

Figure 5. The framework of the Stacking model

B. Modelling

The selection and combination of base classifiers and meta-classifiers are crucial for the Stacking ensemble model. In the selection of base classifiers, this study initially considers seven algorithms as candidates: Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Linear SVM (LSVM), Gaussian Kernel SVM (RSVM), Naive Bayes (NB), and Gradient Boosting Decision Trees (GBDT). The selection of base classifiers follows the principle of "good and diverse," which means choosing models with strong learning abilities and significant differences to improve the overall predictive performance of the model [14]. The accuracy of each classifier is measured to evaluate the learning ability of the model, as shown in Table 2.

TABLE 2. ACCURACY OF EACH CLASSIFIER

Classifier	Accuracy
RF	85.90%
LR	85.57%
KNN	82.62%
LSVM	86.23%
RSVM	84.92%
NB	85.57%
GBDT	81.64%

From Table 2, it can be observed that the accuracy of each classifier is above 80%, indicating a certain level of predictive capability. Among them, the Linear SVM (LSVM) has the highest accuracy, reaching 86.23%, while the Gradient Boosting Decision Trees (GBDT) has the lowest accuracy, only 81.64%. To further consider the diversity of individual classifiers, we compare the Pearson correlation coefficients among the models to measure their differences. The correlation coefficients are calculated based on the class probability values of each classifier on the original training set. The correlation matrix heatmap is shown in Figure 6.

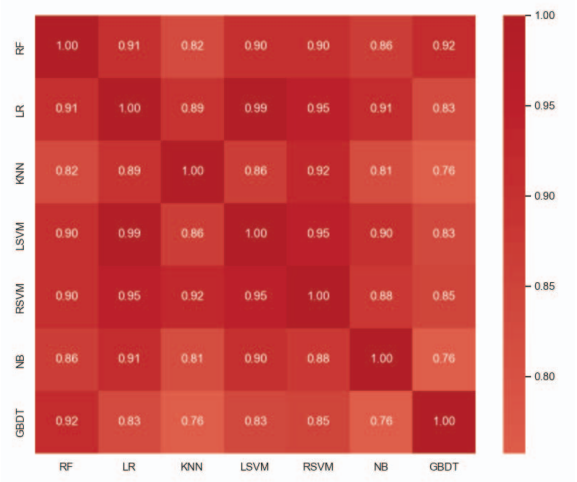

Figure 6. A Heatmap of Correlation Matrix of Base Classifiers

By examining the lower values in Figure 6, which indicate lower correlation coefficients and greater differences between classifiers, we can observe that the correlation coefficients among the classifiers are all above 0.7, indicating strong correlations. The lowest correlation coefficient is between GBDT and NB, as well as between GBDT and KNN, with a value of only 0.76. The correlation coefficient between NB and KNN is 0.81, which is also relatively low. Therefore, based on this observation, we tentatively select these three classifiers (GBDT, NB, and KNN) as the base classifiers for further research.

In the selection of the second-layer meta-classifier, it is generally recommended to choose a model with strong generalization ability and simplicity. This is done to correct any biases introduced by multiple algorithms on the training set and to prevent overfitting [15]. For classification problems, existing researches have shown that logistic regression is a good choice [16]. Therefore, the paper uses logistic regression as the meta-classifier.

Based on the above, the paper selects KNN, NB, and GBDT as the base classifiers, and LR as the meta-classifier. The class probabilities generated by KNN, NB, and GBDT are used to create a secondary training set. The LR algorithm is then trained on this set and tested on the secondary test set. The stacking model using this combination achieves a final accuracy of 88.52%, higher than the accuracy of any individual classifier. To further validate the rationale for the selection of base classifiers, various combinations of base classifiers were trained, and the results are shown in Table 3. It can be observed that their accuracies are all lower than 88.52%.

TABLE 3. DIFFERENT COMBINATIONS OF BASE CLASSIFIERS AND THEIR ACCURACY

Combination	Accuracy
KNN+NB+GBDT	88.52%
LR+LSVM+RSVM	83.61%
RF+KNN	83.61%
RF+KNN+LR	86.89%
LR+LSVM+GBDT	85.25%
RSVM+LSVM+GBDT	83.61%

Finally, the heart disease classification prediction model based on stacking is obtained, and its specific framework is shown in Figure 7.

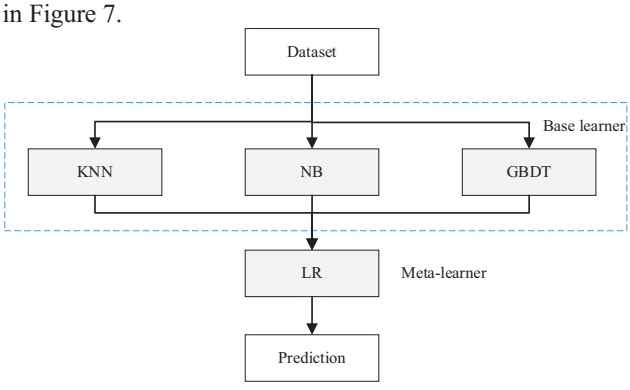

Figure 7. Our model frame diagram

C. Evaluation Metrics

Due to the issue of imbalanced data distribution, the classification results are often influenced by the majority class, leading to a bias towards the class with more samples and overlooking the class with fewer samples. Using a single evaluation metric makes it challenging to comprehensively assess the model's performance. Therefore, this study utilizes the comprehensive evaluation metrics of the ROC curve and AUC to provide a more accurate and comprehensive evaluation of the model. The ROC curve plots the true positive rate against the false positive rate for various classification thresholds, and the AUC represents the area under the ROC curve. These metrics allow for a thorough assessment of the model's performance. Please refer to Figure 8 for the visualization of the ROC curve and AUC.

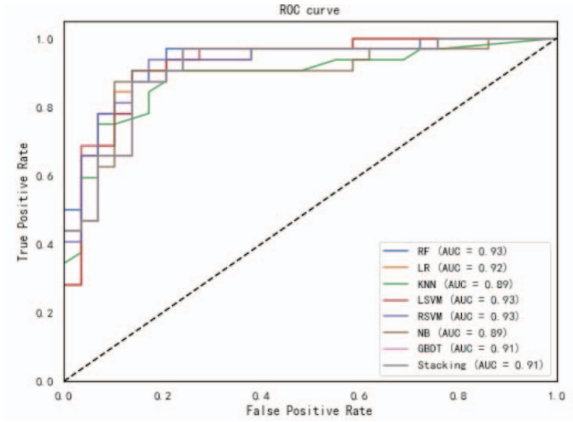

Figure 8. ROC curves of models

It can be observed that the ROC curves of each classifier are relatively close to each other, without one curve completely enclosing another. The visual comparability is not strong. By calculating the area under the ROC curve (AUC) for each classifier, it is found that the AUC values for all classifiers fall within the range of 0.85 to 0.95, indicating good performance. Among them, the AUC of the Stacking classifier is 0.91, which is lower than RF, LR, LSVM, and RSVM. This suggests that the Stacking model still has room for improvement and further optimization.

IV. COST-SENSITIVE LEARNING

A. Basic conception

Traditional machine learning methods always assume equal class sizes or equal misclassification costs, which is clearly unreasonable. This approach can easily lead to misclassifications and degrade classification performance [17]. However, the dataset chosen in this study clearly exhibits such a situation.

Currently, cost-sensitive learning methods can be broadly classified into the following three categories [18]: data preprocessing methods, post-processing methods, and direct cost-sensitive learning methods.

- Data preprocessing methods primarily include sampling methods and weighting methods, which aim to modify the original dataset to create a balanced dataset.

- Post-processing methods address the cost-sensitive learning problem by adjusting the classifier's decision threshold. Examples of such methods include MetaCost, ETA, and cost-sensitive neural networks.
- Direct cost-sensitive learning methods directly embed cost information into the objective function of the learning model. By minimizing the expected loss, these methods obtain the final learning algorithm. The specific algorithms employed in this approach vary depending on different algorithms and application scenarios [19].

B. Threshold Optimization

We further introduce the idea of cost-sensitive learning. For the prediction of heart disease, misdiagnosing healthy people as patients will only increase some follow-up examinations in most cases and waste certain medical resources. However, misdiagnosing patients as healthy people is likely to make patients miss the optimal treatment time and eventually lose their lives. Compared to the former, the latter error is more serious. Therefore, it is necessary to adjust the threshold appropriately to enhance the ability to identify individuals with heart disease. In contrast, the traditional logistic regression model defaults to a threshold of 0.5, classifying values greater than 0.5 as diseased and values less than 0.5 as healthy. Since there is a need to enhance the recognition of individuals with heart disease, the threshold needs to be lowered to increase the probability of identifying individuals as diseased.

Based on the threshold optimization logistic regression model for breast cancer prediction proposed by Mavaddat et al. [20], the paper further optimized the model to determine the optimal classification threshold. During the process of plotting the ROC curve, thresholds are selected one by one. Therefore, in this study, the ROC curve was plotted on the secondary training set, and the F1-score was used as the evaluation metric. The threshold corresponding to the highest F1-score was selected as the optimal classification threshold. The optimal threshold was found to be 0.4019*, as shown in Figure 9. Finally, by adjusting the predictions of the logistic regression model using the optimal threshold on the secondary test set, it was observed that the accuracy was further improved. The final accuracy reached 90.16%, which is a 1.64% improvement compared to the original model.

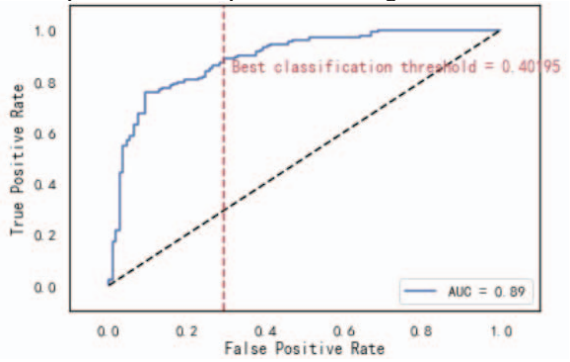

Figure 10. Best classification threshold

V. CONCLUSIONS

The paper focuses on building a stacking model for heart disease data and determining the best combination of base classifiers based on their correlation coefficients. KNN, NB, and GBDT are chosen as the base classifiers in the first layer, and 5-fold cross-validation is applied to prevent overfitting. A simple logistic regression model is used as the meta-classifier in the second layer, and it is found that the final model achieves higher accuracy than individual classifiers, demonstrating good generalization ability.

To address the issue of imbalanced heart disease data, cost-sensitive learning is introduced, and further improvement is achieved by optimizing the threshold to adjust the prediction results. This leads to an increased accuracy of the model, making it more suitable for diagnostic prediction in the medical field. However, most of the parameters in this study are set to their default values, and further research can be conducted to optimize the overall performance of the model through parameter tuning.

ACKNOWLEDGMENT

This work was supported in part by National Social Science Fund General Project of China under Grant 18BTQ058, National Natural Science Foundation of China under Grant 61502299, and the Science and Technology Key Project of Jiangxi Province under Grant 20142BBE50015.

REFERENCES

[1] T. W. C. of the Report, "Report on cardiovascular health and diseases in China 2021: an updated summary," Biomedical and Environmental Sciences, vol. 35, no. 7, pp. 573-603, 2022.

[2] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," Materials Today: Proceedings, vol. 80, pp. 3682-3685, 2023.

[3] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 10-19, 2021.

[4] K. Wisaeng, "Predict the diagnosis of heart disease using feature selection and k-nearest neighbor algorithm," Applied Mathematical Sciences, vol. 8, no. 83, pp. 4103-4113, 2014.

[5] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," International Journal of Pharmaceutical Research, vol. 12, no. 4, pp. 56-66, 2020.

[6] S. Ramachandran, J. French, M. Vanderpump, P. Croft, and R. Neary, "Using the Framingham model to predict heart disease in the United Kingdom: retrospective study," Bmj, vol. 320, no. 7236, pp. 676-677, 2000.

[7] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: a survey," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020: IEEE, pp. 302-305.

[8] M. A. Khan and Y. Kim, "Cardiac Arrhythmia Disease Classification Using LSTM Deep Learning Approach," Computers, Materials & Continua, vol. 67, no. 1, 2021.

[9] Y. Wang, "Identification of Cardiovascular Diseases Based on Machine Learning," in Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences, 2022, pp. 531-536.

[10] X. Guo, Y. Gao, D. Zheng, Y. Ning, and Q. Zhao, "Study on short-term photovoltaic power prediction model based on the Stacking ensemble learning," Energy Reports, vol. 6, pp. 1424-1431, 2020.

[11] Sharma, R. and Singh, S.N. “Towards Accurate Heart Disease Prediction System: An Enhanced Machine Learning Approach,” International Journal of Performability Engineering, vol. 18, no. 2, pp. 136, 2022.

[12] D. H. Wolpert, "Stacked generalization," Neural networks, vol. 5, no. 2, pp. 241-259, 1992.

[13] H. Zhang, J.-L. Li, X.-M. Liu, and C. Dong, "Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection," Future Generation Computer Systems, vol. 122, pp. 130-143, 2021.

[14] N. Li et al., "Underground mine truck travel time prediction based on stacking integrated learning," Engineering Applications of Artificial Intelligence, vol. 120, p. 105873, 2023.

[15] X. Ying, "An overview of overfitting and its solutions," in Journal of physics: Conference series, 2019, vol. 1168: IOP Publishing, p. 022022.

[16] X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," in Proceedings of the 22nd international conference on Machine learning, 2005, pp. 505-512.

[17] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to CNN-based image classification," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 4, pp. 1239-1253, 2019.

[18] V. S. Sheng and C. X. Ling, "Thresholding for making classifiers cost-sensitive," in Aaai, 2006, vol. 6, pp. 476-481.

[19] M. Torabbeigi, G. J. Lim, N. Ahmadian, and S. J. Kim, "An optimization approach to minimize the expected loss of demand considering drone failures in drone delivery scheduling," Journal of Intelligent & Robotic Systems, vol. 102, no. 1, p. 22, 2021.

[20] N. Mavaddat et al., "Polygenic risk scores for prediction of breast cancer and breast cancer subtypes," The American Journal of Human Genetics, vol. 104, no. 1, pp. 21-34, 2019.