Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240641

Cardiovascular disease prediction based on Stacking integrated strategy

Jialiang Sun

College of medical Instruments, Shenyang Pharmaceutical University, Shenyang, China

s13893631739@ldy.edu.rs

Abstract. According to the report of the World Health Organization, cardiovascular disease causes at least 17.1 million deaths worldwide every year, ranking second among the top ten causes of death for many years, and is still a problem to be solved in China and even in the world. No less dangerous than cancer on the list. Early diagnosis is of great significance for patients with cardiovascular diseases, which can diagnose patients with cardiovascular diseases as early as possible to achieve the purpose of early treatment and reduce the cost and pain of patients. In order to achieve accurate early diagnosis of cardiovascular diseases, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting (GBDT), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), Deep Neural Network (DNN), and Stacking integrated models were used to predict cardiovascular diseases. The comparison results showed that the Stacking model is the optimal prediction model. The precision reached 86.80%, the recall reached 84.78%, and the f1 reached 85.76%. The proposed model can be used in cardiovascular disease prediction to reduce the incidence of cardiovascular disease.

Keywords: Machine learning, Cardiovascular disease, Stacking, forecasting

1. Introduction

According to the latest data from the World Health Organization (WHO), cardiovascular disease (CVD) continues to be the foremost cause of mortality globally, presenting a formidable challenge to human health. As such, the exploration and prediction of CVD represent a common imperative for mankind [1]. In the context of China, the acceleration of aging demographics, coupled with prevalent unhealthy lifestyles and habits, has engendered a substantial population susceptible to cardiovascular risk factors, thereby exacerbating the burden of CVD in the nation.The prevalence of CVD in China is alarming, with approximately 330 million individuals afflicted by various cardiovascular ailments, encompassing conditions such as stroke, coronary heart disease, heart failure, pulmonary heart disease, atrial fibrillation, rheumatic heart disease, congenital heart disease, peripheral artery disease, and hypertension. The economic ramifications are significant, with the total inpatient expenditure for cardiovascular and cerebrovascular diseases in 2020 reaching a staggering 270.901 billion yuan [2].Efforts aimed at averting and managing CVD are paramount. Given the high morbidity, mortality rates, and therapeutic challenges associated with these diseases, early detection and intervention stand as the cornerstone of effective management. However, conventional approaches to disease detection and prognostication often hinge upon the subjective discernment of healthcare professionals.To bolster the objectivity and

DOI: 10.54254/2755-2721/67/20240641

precision of predictive analytics, the integration of machine learning (ML) methodologies has emerged as a pivotal paradigm. By synthesizing clinical data with advanced ML algorithms, healthcare practitioners can glean actionable insights for more nuanced and accurate disease prognostication, thereby facilitating prompt interventions and treatment modalities for patients.

2. Literature Review

Feature attributes required for model construction were gathered by Jiang et al. from both the UCI database and the Kaggle platform. A total of 27 features and 1024 cases were screened for training and testing purposes. In their study, a Long Short-Term Memory (LSTM) neural network model was utilized for cardiovascular disease prediction. Notably, their approach surpassed traditional models such as support vector machines and K-nearest neighbor models, thereby addressing a research gap by directly applying LSTM models for cardiovascular disease prediction. However, the small sample size selected resulted in instability and bias [3].Data from the Kaggle platform was leveraged by Fang et al., who employed the CatBoost model to mitigate overfitting and enhance prediction capability for cardiovascular diseases. Their model exhibited significantly improved prediction performance compared to traditional approaches [4].Ke et al. devised a cardiovascular disease prediction model utilizing LightGBM and optimized it with the K-nearest neighbor algorithm. The efficiency and accuracy of LightGBM proved effective in handling large-scale data and complex features [5].Li established a neural network-based cardiovascular disease prediction model, yielding promising results. However, the small sample size of only 1354 individuals may introduce bias and instability [6].Comprehensive big data analysis was conducted by Wang et al. using Hive, resulting in the establishment of a high-quality visual analysis framework for cardiovascular diseases. Their approach aimed to enhance data precision and specificity, facilitating clearer data classification for different user groups and providing tailored insights and references [7].In summary, while numerous studies have focused on predicting cardiovascular disease using machine learning models and most of the models used are relatively simple and unable to overcome inherent limitations. To address the diagnosis of cardiovascular diseases and improve the accuracy of early diagnosis, an integrated model based on Stacking is proposed. The performance of this model was compared with that of logistic regression, decision trees, random forests, gradient boosting decision trees (GBDT), AdaBoost, XGBoost, and deep neural networks (DNN).

3. Method

3.1. Cardiovascular Disease dataset

The current research data utilized originates from Kaggle [8], comprising cardiovascular disease data encompassing a substantial sample size of 70,000 subjects. This dataset is characterized by 11 key attributes: Age, Height, Weight, Gender, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose, Smoking, Alcohol intake, and Physical activity.The sample population is comprised of 35% men and 65% women, with the majority falling within the age range of 40 to 65 years. Among the subjects, patients make up approximately 49.97% of the sample, while non-patients account for the remaining 50.03%.It's worth noting that the dataset is meticulously balanced, with no missing values, ensuring a comprehensive and reliable foundation for the research analysis and predictions.

3.2. Descriptive statistics

3.2.1. Age, height,weight and gender were associated with cardiovascular disease

Through the analysis of this data set, it is found that age has a huge impact on cardiovascular diseases, and the age of patients is generally over 55 years old, indicating that the average age of disease is generally high. Therefore, it can be concluded that if people are over 55 years old, they should pay attention to the prevention of cardiovascular diseases and take regular physical examinations. Avoid the harm of cardiovascular disease to physical health and its damage to property. Weight also has an impact

DOI: 10.54254/2755-2721/67/20240641

on cardiovascular disease, with the weight of people with the disease generally higher than that of people without the disease. This means that the higher the weight, the higher the risk of cardiovascular disease. Therefore, people of all ages should pay attention to exercise and maintain a healthy weight to prevent cardiovascular disease. Gender and height had no effect on the incidence of cardiovascular disease. However, based on the data set, the male proportion is about 6.5 percent, and the gender ratio is unbalanced, which may affect the results.

3.2.2. Cholesterol, blood sugar, Systolic blood pressure, diastolic blood pressure and cardiovascular disease

Through the analysis of the data, it was found that cholesterol and blood sugar had a greater impact on cardiovascular disease, and the proportion of people with elevated cholesterol borderline was higher than that of the normal population. So high or low cholesterol can lead to an increased risk of cardiovascular disease. The proportion of cardiovascular diseases in people with elevated blood glucose margin is also higher than that in normal people, indicating that cardiovascular and blood glucose have similar effects on cardiovascular diseases, and too high or too low blood glucose will also lead to an increase in the prevalence of cardiovascular diseases. And it is worth noting that high cholesterol or high blood sugar is more likely to cause cardiovascular disease. So in life, people should pay attention to diet and lifestyle. To prevent cardiovascular disease. Systolic and diastolic blood pressure also have an impact on cardiovascular disease, and people with high systolic and diastolic blood pressure are more likely to have cardiovascular disease. And if a person is not having CVD, then There's more likely (55.3 %) that he/she has 120 mmHg Daistolic Blood Pressure, If a person is having CVD, then There's more likely (42.5 %) that he/she has 120 mmHg Systolic Blood Pressure with second mostly likely case (31.9%) of having 90mmHg Daistolic Blood Pressure.

3.2.3. Smoking, alcohol consumption, regular exercise and cardiovascular disease

Smoking and alcohol consumption were not identified as significant factors influencing the risk of cardiovascular disease. However, individuals who engage in regular exercise exhibit fewer cardiovascular issues compared to those who lead a sedentary lifestyle. This highlights the preventive benefits of regular physical activity against cardiovascular disease. Given that people often lead sedentary lifestyles, coupled with busy schedules and a preference for high-calorie foods, the risk of cardiovascular disease tends to increase. Establishing healthy exercise habits is evidently crucial in mitigating this risk and promoting cardiovascular health. Overall, lifestyle habits can greatly affect the probability of cardiovascular disease. So people should pay more attention to their lifestyle. A healthy lifestyle can effectively prevent cardiovascular disease.

3.3. Feature Screening

In this study, feature selection is conducted using the decision tree algorithm, a widely employed method in machine learning. Decision trees determine feature importance and relevance by iteratively splitting nodes based on the chosen features. At each node, the decision tree evaluates various criteria to select the best feature that maximizes information gain or minimizes the reduction of Gini impurity. This iterative process continues until a stopping criterion is met, resulting in the creation of a tree structure where each leaf node represents a class label or regression value.

By strategically splitting nodes and selecting appropriate features, decision trees can effectively partition the dataset into subsets with higher purity, enhancing the model's ability to perform classification or regression tasks accurately. Common criteria for selecting the best features include information gain, information gain ratio, and Gini impurity.

After applying the decision tree algorithm for feature selection, 10 out of the initial 11 features are retained for further analysis. These features include age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol level, glucose level, smoking status, and physical activity level. The screening process ensures that only the most relevant and informative features are included in the predictive model, optimizing its performance and interpretability (Figure 1 and Figure 2).

DOI: 10.54254/2755-2721/67/20240641

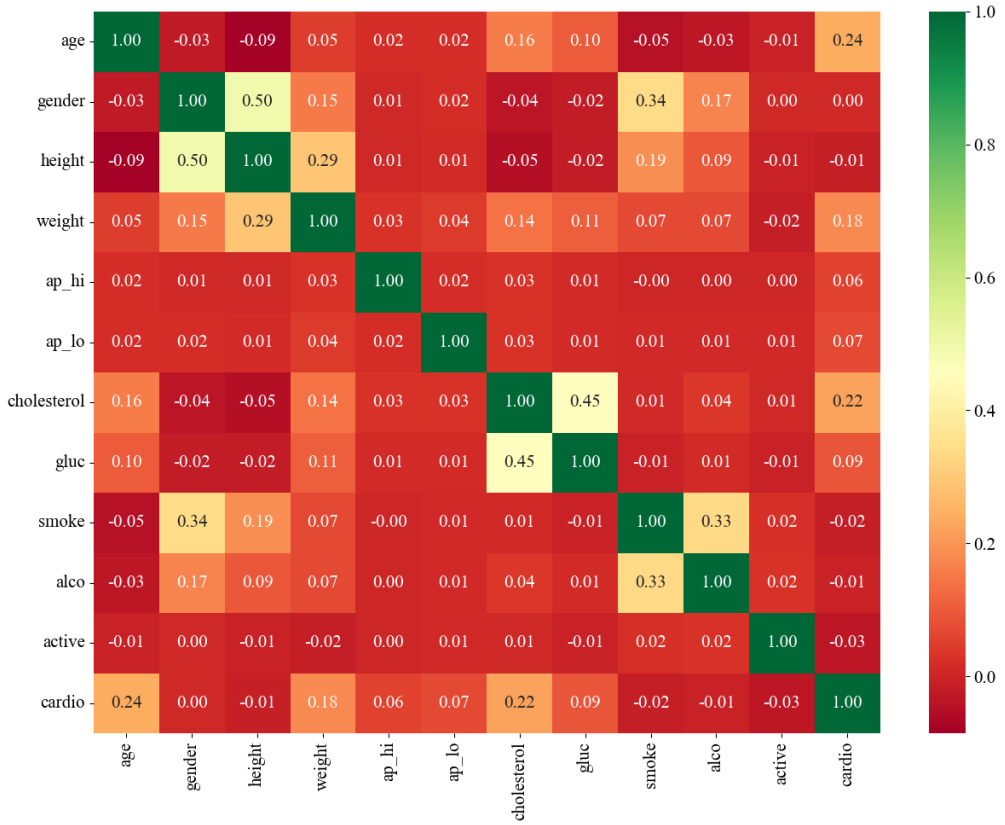

Figure 1. Correlation matrix of features.

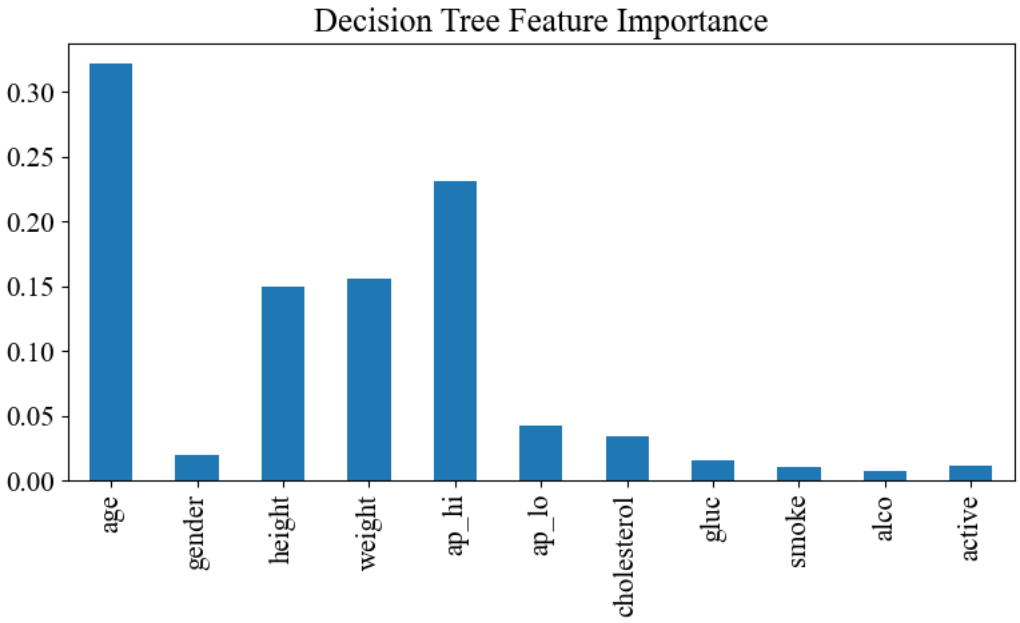

Figure 2. Decision Tree Feature Importance

DOI: 10.54254/2755-2721/67/20240641

3.4. Model structure and implementation

GBDT, KNN, XGBoost, AdaBoost and Logistic regression are several models commonly used for disease prediction. GBDT is capable of handling various types of data, including both continuous and discrete features, and does not require much feature engineering when working with high-dimensional data. GBDT is capable of dealing with nonlinear relationships, insensitive to outliers, and has good generalization ability. However, GBDT is prone to overfitting on high-noise data sets and is not efficient for large-scale data processing. KNN is easy to understand, requires no training process, and learns based on examples. However, for high-dimensional sparse data sets, its performance is poor. XGBoost overcomes KNN's inefficiencies in large-scale data processing, but is sensitive to outliers. AdaBoost is not easy to overfit, but has limited ability to handle nonlinear relationships. Logistic regression is easy to understand, but it is poor at modeling nonlinear relationships.

Each of these models has advantages and disadvantages in predicting cardiovascular disease. This study considers the integrated Stacking algorithm to combine them into a unified model.

Stacking is an ensemble learning technology that combines multiple models to improve forecasting performance. The principle is: first train a different set of base models on the training data, and after the training is completed, use each base model to make predictions on the validation set. The predictions of the base model are then used as features to train a metamodel that learns how to combine the predictions of the base model and make a final prediction. Because different basic models are allowed to be used in Stacking, the diversity of models is greatly enhanced, and it has good generalization ability. Its flexibility also enables it to adapt to various environments

The Stacking structure consists of two layers of algorithms in sequence. The first layer (level 0) contains at least one strong learner and the second layer (level 1) can accommodate only one learner. In the training process, the data is input to the first layer for training, and each algorithm will output the corresponding prediction results. These results are then spliced into a new feature matrix, which is then fed into a second-layer algorithm for further training. The final output of the fusion model is determined by the second layer learner. In this paper, the basic models of the first layer are GBDT, KNN, XGBoost, and AdaBoost, while the meta-model of the second layer is Logistic regression (Figure 3).

In the evaluation of the performance of the second layer meta-model (LR), the feature data set generated by the first layer model is cross-validated by 10 fold. The training data is divided into 10 subsets of similar size, with each subset serving as the training set and the rest as the validation set. This process is repeated ten times, and the average of the ten validation results is used as the final performance evaluation indicator (Figure 4).

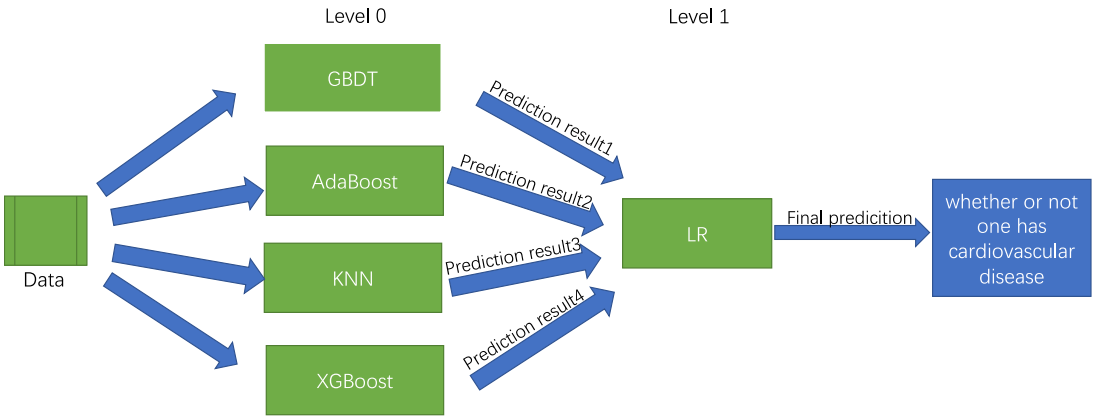

Figure 3. Model architecture

DOI: 10.54254/2755-2721/67/20240641

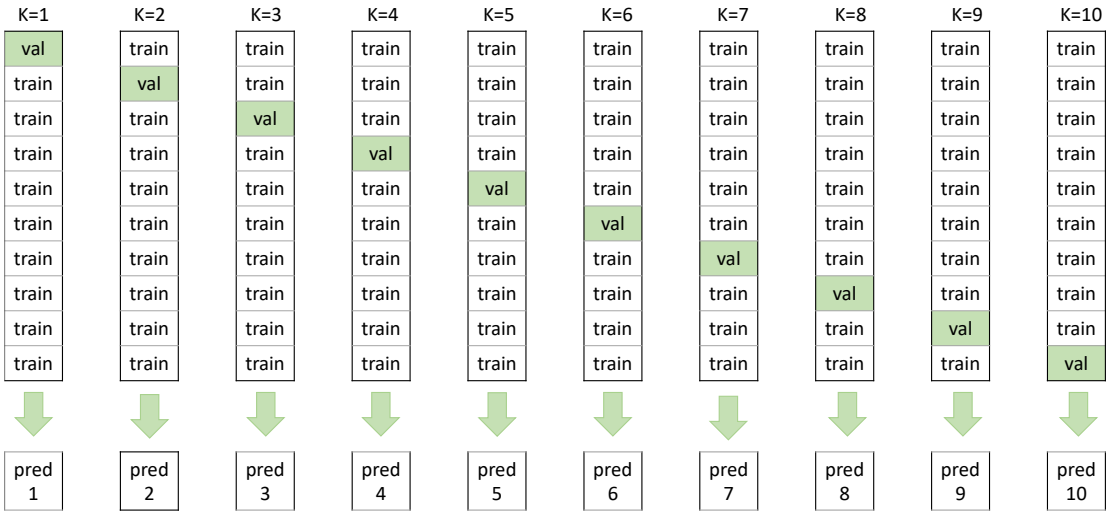

Figure 4. Ten-fold cross-validation

3.5. Evaluation metrics

In this study, the proposed method is evaluated on a number of indicators. For the binary classification learning task, the actual category of the sample to be tested and the category predicted by the classifier can be divided into true examples, false positive examples, true counter examples and false counter examples, and their corresponding numbers are represented by TP, FP, TN and FN respectively, then the total sample number $F = TP + FP + TN + FN$

Accuracy refers to the ratio of the number of samples correctly classified by the classification model on a given data set to the total number of samples. Accuracy is important, however, when the number of samples in different categories is unbalanced, accuracy may not be a good indicator of performance. Because the model may tend to predict categories with a larger sample size. The recall rate is a measure of the ability of a classification model to identify positive samples. The recall rate represents the proportion of all samples that are actually positive that the model successfully predicts to be positive. The recall rate ranges from 0 to 1, with a value closer to 1 indicating better performance of the model in identifying positive samples. If the recall rate is high, it means that the model has successfully captured most of the positive samples, but it may lead to an increase in the number of false positive examples. Recall rate is the most important indicator in this study. The F1 score is an indicator that comprehensively evaluates the performance of a classification model, taking into account the accuracy and recall rate of the model. The F1 score is the harmonic average of accuracy and recall. F1 scores range from 0 to 1, with a value closer to 1 indicating better model performance. F1 scores are suitable for a variety of category imbalances and are commonly used when evaluating the performance of a classification model. It synthesizes the accuracy and recall rate of the model, considering not only the ability of the model to identify positive samples, but also the accuracy of the model in predicting the positive class. In some application scenarios, an F1 score may be a better choice. Accuracy is an important index to measure the performance of a classification model, which represents the ratio of the number of samples correctly classified by the model on the whole data set to the total number of samples. Accuracy ranges from 0 to 1, and the closer the value is to 1, the higher the prediction accuracy of the model. Accuracy is one of the most commonly used metrics to evaluate the performance of a classification model, but it can be insufficiently comprehensive when dealing with class-unbalanced datasets.

The accuracy rate P and recall rate R, f1, Acc are calculated by the following formula.

DOI: 10.54254/2755-2721/67/20240641

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

$$\text{Acc}uracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

4. Results

4.1. Model parameters

The importance of recall in disease prediction is self-evident. Because false positive samples can lead to the disease not being diagnosed in time, which can cause serious damage to the patient's health. Assuming that a disease prediction system has a low recall rate means that the system may miss some cases with the disease, which will lead to missed opportunities for early intervention and treatment for these patients. If the recall rate is higher, the system can more comprehensively identify the cases of the disease, so as to take the necessary medical measures in time to protect the health and life safety of patients to the greatest extent. Therefore, in the prediction of diseases, the recall rate should be increased as much as possible to ensure that as many patients as possible get timely diagnosis and treatment, so as to minimize the harm caused by diseases to patients.In this study, grid search was conducted, with Recall as the index, and then the parameter range was defined to search for the best parameter combination within the range. Finally, the LR regularization parameter was set to 0.1, the class weight was set to balanced, max iteration was set to 100, and solver was set to newton-cg.

4.2. Results description

To verify the superiority of the Stacking model, two test sets were used to compare it with several other models, including random forests, XGBoost, logistic regression, decision trees, DNN, and Adaboost. The comparison results are shown below. It can be seen from the table that the Stacking model performs best on multiple indicators such as accuracy (Acc), Recall, and F1 score, while its Precision is only 0.18% lower than that of Adaboost. Therefore, the Stacking model performs best among these models. These experimental results fully demonstrate the validity and robustness of the Stacking model in predicting target diseases and validate its superiority over other models tested.

One-way ANOVA of variance was used in this paper. The Acc (p=0.99), Precision (p=0.99), Recall (p=1.0) and F1 (p=1.0) showed no statistical difference. Although there is no statistical difference among these groups, the mean value of stacking groups is higher than that of other stacking groups, indicating the best performance of stacking groups (Figure 5).

Table 1. Model comparison evaluation index

	LR (%)	DT (%)	RF (%)	GBDT (%)	AdaBoost (%)	XGBoost (%)	DNN (%)	Stacking (%)
Acc	80.19± 21.41	79.95± 23.16	84.30± 18.43	85.26± 16.61	85.38± 17.75	85.78± 18.24	77.44± 18.24	86.04± 16.92
Precision	79.26± 18.61	80.35± 23.70	84.70± 18.38	86.08± 15.64	86.98± 16.33	86.74± 17.30	77.82± 17.30	86.80± 16.35
Recall	77.99± 23.00	79.98± 23.44	83.95± 19.45	84.07± 19.28	82.19± 20.90	84.08± 20.90	79.50± 20.05	84.78± 19.21
F1	79.87± 22.67	79.70± 22.92	84.32± 18.92	85.18± 17.71	84.52± 18.69	85.34± 18.69	78.64± 18.65	85.76± 17.82

DOI: 10.54254/2755-2721/67/20240641

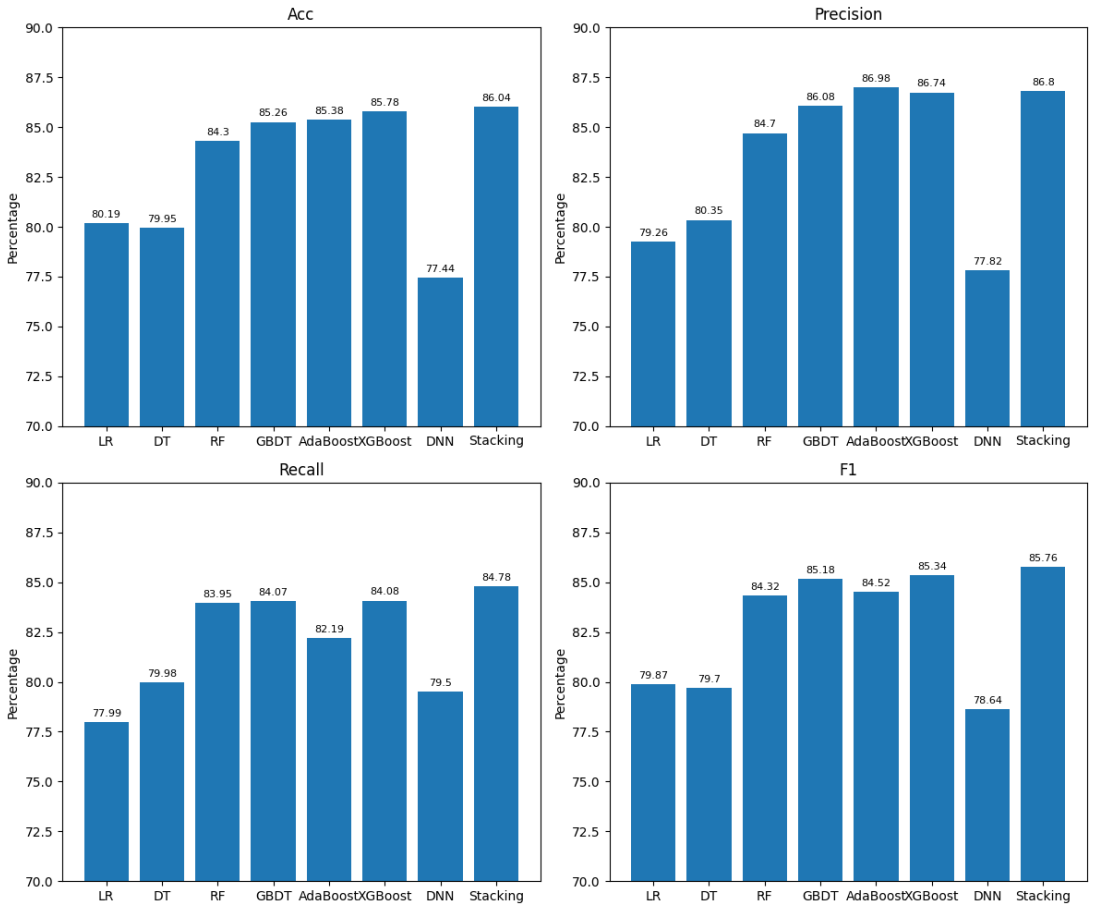

Figure 5. The results of model comparison.

5. Discussion

Various studies have focused on cardiovascular disease prediction, however, most of them only applied single classical machine learning model which may not generate an accurate and realibale result when it encounters complex clinical data. In the present study, a stacking integrated strategy was proposed to solve this problem. Due to its ability to integrate multiple basic models, the Stacking model leverages the strengths of different models to offset the weaknesses of individual ones. Its inherent flexibility allows for the combination of diverse model types, contributing to its robustness. These attributes render the Stacking model well-suited for cardiovascular disease prediction. In this study, Gradient Boosting Decision Trees (GBDT), AdaBoost, K-Nearest Neighbors (KNN), and XGBoost were employed as base models, with Logistic Regression (LR) serving as the meta-model to construct the Stacking ensemble. Feature selection was conducted using decision trees, while LR parameters were fine-tuned via grid search. Ultimately, a predictive model for cardiovascular diseases was developed. Models such as LR, DT, RF, GBDT, AdaBoost, XGBoost, and DNN may have some limitations when used alone. LR usually assumes a linear relationship between features and may not perform well in complex data sets. DT and RF are prone to overfitting, especially when dealing with high-dimensional data. GBDT and AdaBoost may have longer training time and are more sensitive to outliers. XGBoost and DNN generally require more hyperparameter tuning and compute resources. In contrast, the Stacking model can synthesize the advantages of basic models and improve forecasting performance through metamodel learning. It can effectively reduce the limitations of individual models and realize effective combinations between different models, thus improving the overall prediction accuracy. This method of synthesizing multiple models enables the Stacking model to have significant advantages in complex forecasting tasks.

DOI: 10.54254/2755-2721/67/20240641

Besides the Stacking model's advantages, it's worth mentioning the strengths of the individual models employed in the ensemble. LR offers simplicity and interpretability, while DT and RF excel at capturing complex relationships in the data. GBDT and AdaBoost provide robustness and can handle large datasets, while XGBoost and DNN offer high predictive accuracy and can capture intricate patterns in the data. By combining these models, the Stacking ensemble leverages their diverse strengths to enhance predictive performance.

Despite its advantages, the study also has several limitations. Firstly, the model used in this study has a relatively simple structure and may not capture the complexity of cardiovascular diseases fully. Secondly, the dataset used in the study may be insufficient, which could impact the model's predictive performance. Additionally, the feature set is limited, which may result in important predictive variables being overlooked. Furthermore, the study relies solely on textual data, which may not capture all relevant information for accurate disease prediction.

Moreover, the data cleaning process may not have been thorough enough, potentially introducing noise or bias into the model. Additionally, the parameter tuning process may not have been optimized to its fullest extent, which could affect the model's performance. These shortcomings highlight the need for further research and refinement to enhance the accuracy and reliability of the cardiovascular disease prediction model.

6. Conclusion

Given the challenging nature and high mortality rates associated with cardiovascular diseases, this study employs the Stacking model for prediction, yielding a precision of 86.80% and a recall of 84.78%. The superior performance of the Stacking model compared to other machine learning models underscores its potential significance for diagnosing cardiovascular diseases in the future. As such, leveraging the Stacking model holds promise for enhancing diagnostic accuracy and improving patient outcomes in cardiovascular healthcare.

References

[1] Yang J, Zhang Y, Ma T, et al 2024 Prevalence, disease burden and prediction of cardiovascular diseases in China, 1990-2019 Chinese General Practice 27 233-244

[2] Wang Z, Ma L, Liu M, Fan J and Hu S 2023 China Cardiovascular Health and Disease Report 2022 Chinese Medical Journal 136 2899-2908

[3] Jiang X 2024 Research on Cardiovascular Disease Prediction Based on Deep Learning Algorithms (Shenyang Normal University)

[4] Ding L and Luo P 2017 Research on Default risk warning of P2P online loans based on the integrated Stacking strategy Review of Investment Studies 36 41-54

[5] Fang T 2023 Research on Machine Learning Methods for Cardiovascular Disease Prediction (Heilongjiang University)

[6] Ke Y and Chen K 2022 Research on cardiovascular disease prediction model based on LightGBM Information & Computer 34 71-3

[7] Wang Z, Qin Y, Wang X, et al 2023 A Hive-based big data analysis of cardiovascular diseases China Science and Technology Information 23 87-91

[8] Svetlana U 2019 Cardiovascular Disease dataset Retrieved from (https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset)