DOI: 10.23977/csoc.2022.020104
ISSN 2516-399X Vol. 2 Num. 1

Classification and Prediction of Cardiovascular Patients Based on Optimal Random Forest Algorithm

Jiaqi Huang[1,*], Mingguang Li[2]

[1]School of Artificial Intelligence, Wuchang University of Technology, Wuhan, 430000, China
[2]School of Computing, Nantong University of Science and Technology, Jiangsu, 210000, China
**Corresponding author: h1963250759@163.com*

Keywords: Cardiovascular disease, integrated learning, optimal random forest, machine learning

Abstract: Cardiovascular disease is a high-risk disease and therefore machine learning is needed to classify and predict it in order to aid research in the medical field. A prediction model for classifying cardiovascular patients based on an optimised random forest algorithm and comparing the prediction performance of each model. Using publicly available data on cardiovascular disease from the Kaggle platform, classification prediction models for cardiovascular disease were developed based on an integrated learning approach using Random Forest, Parsimonious Bayes, SVM and AdaBoost algorithms based on 12 indicators that may have an impact on the mortality of patients with cardiovascular disease. and classification prediction effects. Using the multiple averaging method to ensure the accuracy of the algorithms, the four types of AUROC values were observed and visualisation using matlab's powerful toolbox yielded the best ROC curve fit for random forest with an AUC value of 0.90.

1. Introduction

Cardiovascular disease is the number one cause of death worldwide, claiming an estimated 17.9 million lives each year and accounting for 31% of deaths worldwide. Patients with cardiovascular disease or those at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or pre-existing conditions) need early detection and management, and machine learning models can be of great help.

With the rapid development of artificial intelligence and machine learning, traditional medical theoretical analysis, simple statistics and single-factor analysis no longer meet the needs of the medical field for in-depth exploration. Currently, the main approaches to research in the medical field are KNN, support vector machines, plain Bayes, decision trees, neural networks and the traditional random forest.

A great deal of research in machine learning has been applied to the medical field, with decision trees and random forests being the most common. Celestine Iwendi [1] proposed a random forest model based on the AdaBoost algorithm augmented to classify patients, with 94% accuracy in the model. Erwin Cornelius [2] proposed the so-called clustered random forest method to predict the mortality of patients with COVID19. mortality in patients with COVID19. MICHAEL M.

WARD [3] predicts mortality in patients with SLE using a random forest algorithm. Nazmun Nahar [4] predicts whether a patient has liver disease using an integrated algorithm based on J48, LMT, random forest, random tree, REPTree, decision stump and Hoefting tree. Jabar Akhil [5] proposed a random forest prediction model incorporating genetic algorithms to improve the accuracy of the model, which can be successfully used by healthcare professionals to predict heart disease. Lin Yu [6] developed an integrated learning ICU prediction model based on Random Forest, AdaBoost, GDBT and compared it with a logistic model to predict the risk of readmission to ICU for seriously ill patients. Heng-Li Will [7] used ID3, Pandas and Sklearn algorithms in decision trees to classify and predict patient illness. Yang Li [8] used integrated learning algorithms to predict the prevalence of cardiovascular disease (CVD) in patients after testing multiple regression models, classification and regression trees (CART), plain Bayes, bagged trees, AdaBoost and random forests compared to the random forest algorithm with an AUC value of 0.787.

In summary, in order to conduct in-depth research on cardiovascular diseases and improve the detection and prediction results of cardiovascular disease risk, this paper selects 12 pathological symptoms, proposes an integrated learning model consisting of SVM, AdaBoost algorithm, random forest and plain Bayes, applies multiple weak classifiers and integrates the results of multiple decision trees to obtain morbidity and mortality; at the same time, uses lattice-based analysis to find out the optimal parameters of the model to further improve the model accuracy.

2. Acquisition of data and assumptions

2.1. Data sources

The experimental dataset uses a predictive dataset for the classification of cardiovascular patients provided by the Kaggle platform. This dataset can be used to predict mortality in heart failure and contains 12 data evaluation indicators for 299 patients.

2.2. Data pre-processing

In this paper, considering the instability and missing rate of historical data, the data is first pre-processed for data anomalies and missing data. Two main methods are used to deal with data anomalies, the first is to identify values in the range less than 0 as anomalous; the second is to use a box line diagram of the data to identify values that are too large and too small for the data, and to delete the data anomalies as missing values, pending the next step of processing. The data set in this paper is run through the program to see if it is abnormal, and the final data set is obtained without abnormal and missing cases.

2.3. Experimental environment

This paper uses matlab as the programming tool, using the toolbox that comes with matlab. The hardware configuration is 64GB of memory space, 264GB of hard disk space and the software tool version is matlab2021a.

2.4. Experimental procedure

Firstly, this research put the data and the program in the same folder. Secondly, open the four files in the matlab import folder and bring the processed dataset into the model for training, here the training sets used in this study are decision tree algorithm model, support vector machine model, Adaboost model, random forest model and integrated algorithm model, finally the ROC curves of different models are put on the same picture to get the experimental results. The flow chart is shown in Figure 1 below.

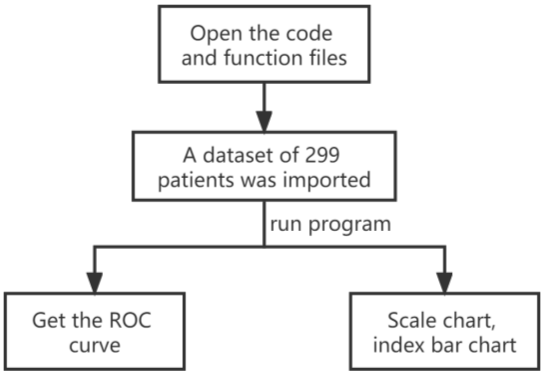
Open the code and function files
A dataset of 299 patients was imported
run program
Get the ROC curve
Scale chart, index bar chart

Figure 1: Flow chart of the experiment

3. A study on the classification and prediction of cardiovascular patients based on an optimized random forest algorithm model

3.1. Theoretical preparation

(1) Integrated learning

Solving a single prediction problem by building a combination of several models. It works by generating multiple classifiers or models that each learn and make predictions independently. These predictions are finally combined into a single prediction and therefore outperform any single classification to make a prediction. The flow chart of how it works was shown in Figure 2 below.

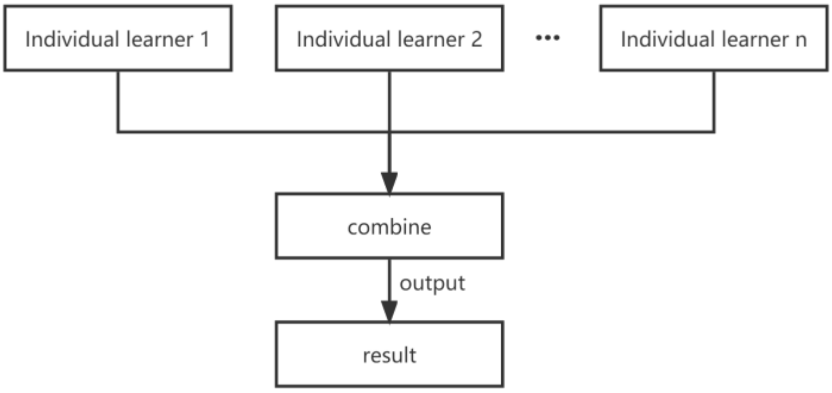

Figure 2: Integration learning flow chart

(2) Decision tree principle

A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. The basic idea is to construct a tree with the fastest decreasing entropy as a measure of information entropy, with zero entropy at the leaf nodes, at which point the instances in each leaf node all belong to the same class. Where the information entropy represents the uncertainty of the information. It is calculated as follows.

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i \mid t)\log_2 p(i \mid t) \tag{1}$$

$p(i \mid t)$ represents the probability that the node t is a classification called i.

Information gain, on the other hand, is a division that can lead to an increase in purity and a decrease in information entropy. It is calculated as follows.

$$Gain(D,a) = Entropy(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} Entropy(D_i) \tag{2}$$

The meaning of equation (2) is the information entropy of the parent node minus the information entropy of all the child nodes. Where D in the equation is the parent node, D_i is the child node, and a in $Gain(D,a)$ is the attribute selection of D i.e. the parent node.

(3) Principle of Random Forest Algorithm

The random forest algorithm is a classification method based on decision trees as learners, and belongs to the Bagging type. It uses CART's decision trees as weak learners, uses bootstrap resampling method to extract multiple samples from the original data, builds a decision tree model based on the samples, and based on the predictions of multiple decision trees, the results of all the classifiers are voted or averaged to the final prediction results are obtained by counting the results of all classifiers by voting or taking the mean. This is shown in Figure 3 below.

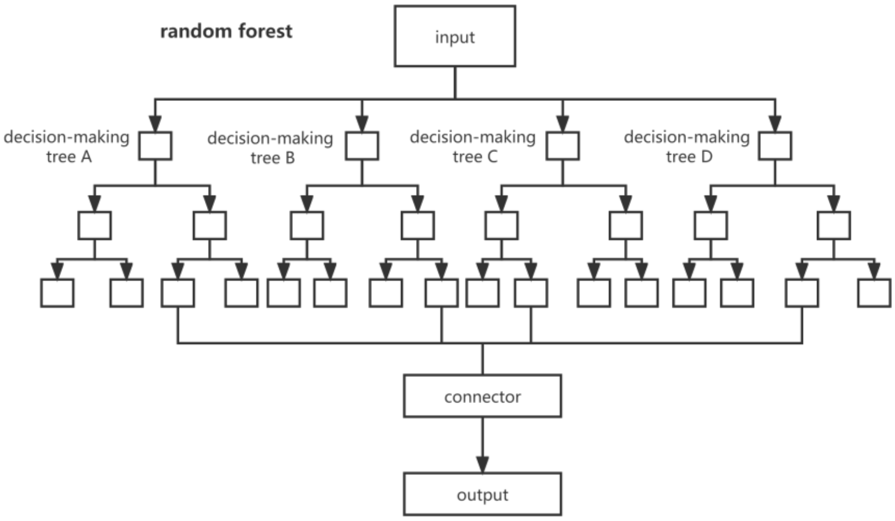

Figure 3: Flow chart of Random Forest algorithm

The CART decision tree used in Random Forest is based on the Gini coefficient for feature selection. The criteria for the selection of the Gini coefficient is that each child node achieves the highest purity, i.e. all observations falling in the child nodes belong to the same classification, at which point the Gini coefficient is smallest, the purity is highest and the uncertainty is smallest.

For a general decision tree, if there are a total of K classes and the probability that a sample belongs to class K is P_K , then the Gini index of this probability distribution is

$$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3}$$

The larger the Gini index, the greater the uncertainty; the smaller the Gini coefficient, the smaller the uncertainty and the more thorough and clean the data split.

For CART trees, because they are binary trees, the following representation can be used.

$$Gini(p) = 2p(1-p) \tag{4}$$

(4) Principle of Support Vector Machine (SVM) algorithm

The support vector machine algorithm is represented in two dimensions by a straight line that divides the dataset on the two-dimensional space into the maximum number of different kinds. If the dataset has a mixed representation on the two-dimensional space and cannot be partitioned, then the space is then up-dimensioned to divide it on the three-dimensional space, and if it still cannot be divided, it continues to be up-dimensioned until it can be divided into different categories.

(5) Principle of AdaBoost algorithm

The AdaBoost algorithm is a boosting algorithm that combines multiple weak classifiers into a strong classifier. It gives the data the same weight when initialising the data for the first training, and then gives the misclassified data a greater weight, i.e. a greater focus on the misclassified data, each time using the results of the previous training as the initial data for training, and repeating this several times to obtain a relatively accurate training This is repeated several times to obtain a relatively accurate training model. The ensemble learning algorithm combines the results of multiple models and selects the best result as the final result of the ensemble learning algorithm.

3.2. Research Methodology

In this paper, the training and test sets were divided into 75% and 25%, respectively, and a total of 12 indicators of cardiovascular disease were selected, namely age, whether the red blood cells or haemoglobin were reduced, level of CPK enzyme, whether the patient had diabetes, ejection fraction, whether the patient contained hypertension, number of platelets in the blood, blood creatinine level, serum sodium, gender, whether the patient smoked, and follow-up period. Firstly, the processed dataset of 299 cardiovascular patients was imported into the model for training. The training set models used in this study were decision tree algorithm model, support vector machine model, AdaBoost model, random forest model, and integrated algorithm model. Secondly, the model without tuning of the parameters was imported for rough training, with multiple simulations to prevent randomness and obtain the model with the greatest accuracy, and then this model was subjected to a grid search to find the optimal parameters. The optimised parameters are then brought into the model and trained to give the optimum accuracy. Finally, the ROC curves of the different models are put on the same image to take advantage of the visualisation to give the researcher a clearer understanding of the superiority of this model.

3.3. Analysis of experimental results

Data from 299 patients with cardiovascular disease were imported to obtain mortality rates for men and women with cardiovascular disease, mortality rates for the 40-90 age group with cardiovascular disease, mortality rates for middle-aged and elderly people with cardiovascular disease, and histograms of the importance ratios of the 12 indicators and ROCs using four models: random forest, SVM, AdaBoost, and plain Bayesian curves.

Figure 4 shows the mortality rate for men and women with cardiovascular disease, which is roughly 32% for men and 31% for women, so women have a slightly lower mortality rate than men, but overall the difference in mortality rates between men and women is minimal. Figure 5 shows the mortality rate for the 40-90 age group with cardiovascular disease, which shows that the 60 year olds with cardiovascular disease have the highest mortality rate at 55%, followed by the 50 year olds with cardiovascular disease who have the second highest mortality rate at approximately 48%.

The mortality rate for people with cardiovascular disease is low at no more than 20% and the mortality rate for people with cardiovascular disease at age 90 is the lowest at no more than 10%.

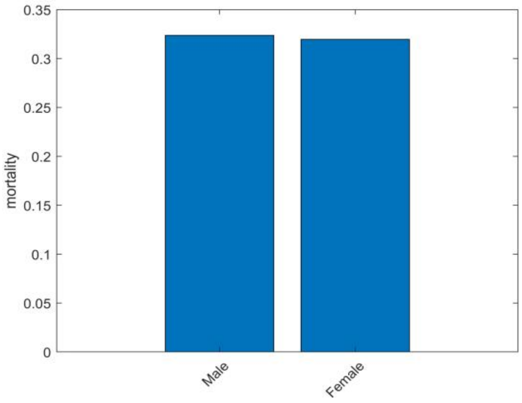

Figure 4: Mortality rates for men and women with cardiovascular disease

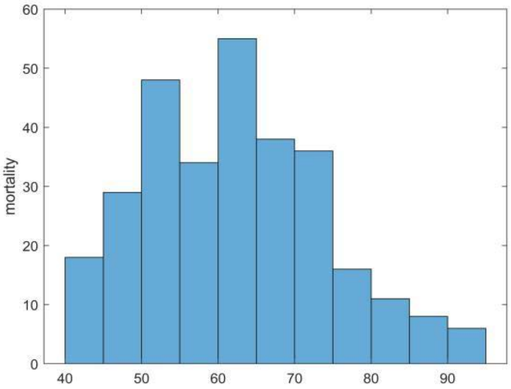

Figure 5: Mortality rates for people aged 40 to 90 years with cardiovascular disease

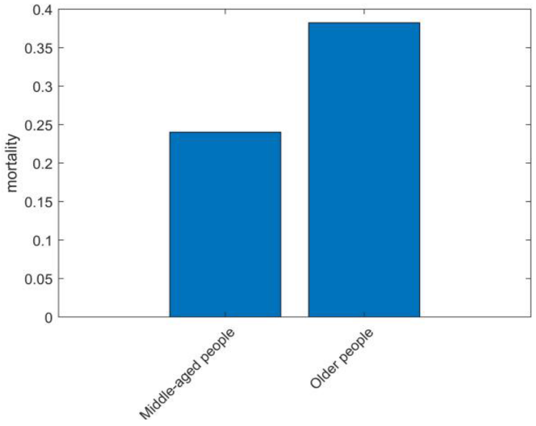

Figure 6: Mortality rates among older adults with cardiovascular disease

As can be seen in Figure 6, the mortality rate for older people with cardiovascular disease is higher than that of middle-aged people, with a mortality rate of roughly 38% for older people compared to roughly 24% for middle-aged people, suggesting that cardiovascular disease is influenced by age, thus producing a greater difference in mortality rates between middle-aged and older people; among deaths, the average age of men is relatively high The average age of men is relatively high among deaths.

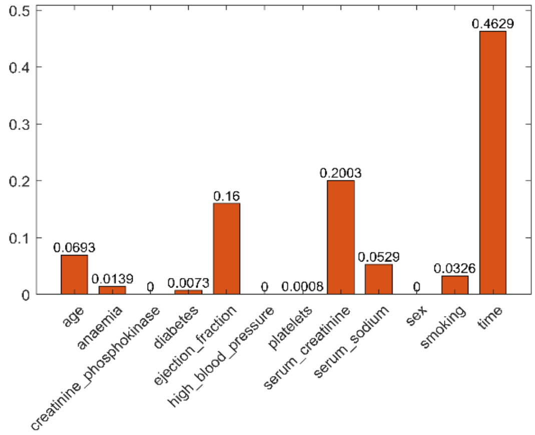

Figure 7: Histogram of the proportion of importance of each indicator

Combining Figures 6 and 7 shows that there are no significant differences in survival status by gender and by the presence or absence of hypertension for both indicators. There was a greater variation in the age distribution of patients suffering from cardiovascular disease, showing a trend towards a lower proportion of survival and a higher proportion of death at older ages. The data show that overall, there is no significant correlation between the presence or absence of smoking and survival.

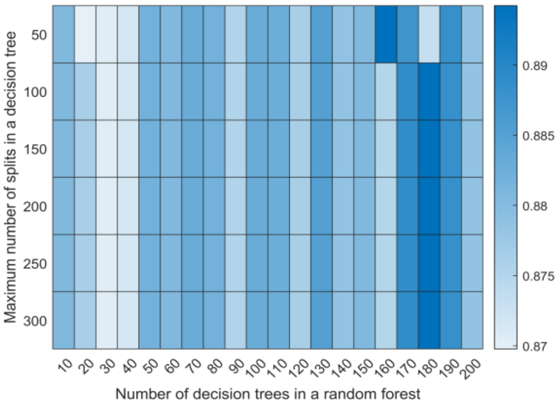

Figure 8: Fractional F1 heat map

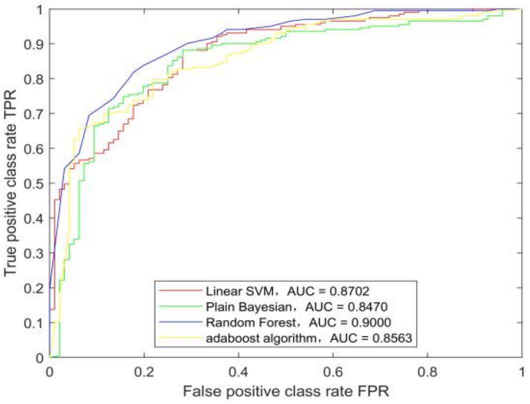

Figure 9: ROC curves and AUC values of the four models

Figure 8 shows a heat map of F1 scores, using grid search for parameter tuning, with the optimization criterion F1. darker grid colours in the figure represent higher scores. Figure 9 shows the comparison of the performance results of the four models on the test set. The AUC value of linear SVM is 0.87, the AUC value of plain Bayes is 0.85, the AUC value of random forest is 0.90, and the AUC value of AdaBoost algorithm is 0.86, among which the AUC value of random forest is the highest, so the ROC curve of random forest has the best result among the four types of models.

4. Conclusion

Most cardiovascular diseases can be prevented by adopting population-wide strategies that address behavioural risk factors such as smoking, unhealthy diet and obesity, lack of physical activity and harmful use of alcohol.

This paper uses an optimised random forest model to investigate the topic of mortality prediction in cardiovascular patients, using the powerful toolbox programming that comes with matlab to visualise and analyse the data, drawing histograms of mortality by gender and age, proportional importance charts for 12 categories of indicators, and ROC curves and AUC values for the four models. Data pre-processing facilitates the execution of algorithms while ensuring sample balance and avoiding one-sided algorithm execution. The integrated learning model is used to combine the results of multiple model training and then select the optimal result as the final result of the ensemble learning algorithm, and then use the gridding analysis to find the optimal parameters of the model to further improve the model accuracy. According to the data, the random forest model has the best fit of the ROC curve with an AUC value of 0.90. As algorithms such as random forest are random in nature, the final result is averaged over multiple results to ensure the accuracy of the algorithm, so the prediction method based on optimised random forest has higher accuracy and better robustness than other methods.

References

[1] Iwendi, Celestine, et al. "COVID-19 patient health prediction using boosted random forest algorithm." Frontiers in public health 8 (2020): 357.

[2] Cornelius, Erwin, Olcay Akman, and Dan Hrozencik. "COVID-19 mortality prediction using machine learning-integrated Random Forest algorithm under varying patient frailty." Mathematics 9.17 (2021): 2043.

[3] Ward, Michael M., et al. "Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of Arthritis Care & Research: Official Journal of the American College of Rheumatology 55.1 (2006): 74-80.

[4] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques. "International Journal of Data Mining & Knowledge Management Process 8.2 (2018): 01-09.

[5] Jabbar, M. A., and Shirina Samreen. "Heart disease prediction system based on hidden naïve bayes classifier. "2016 International Conference on Circuits, Controls, Communications and Computing (I4C). IEEE, 2016.

[6] Lin Y, Wu JY, Lin K, Hu YH, Kong KWL. Predicting the risk of readmission of critically ill patients to intensive care units based on integrated learning models [J]. Journal of Peking University (Medical Edition),2021,53(03):566-572.

[7] Jiang Xingli, Wang Jianhui. Disease classification prediction based on decision tree algorithm [J]. Information and Computers (Theoretical Edition), 2021, 33 (11): 51-53.

[8] Yang, Li, et al. "Study of cardiovascular disease prediction model based on random forest in eastern China." Scientific reports 10.1 (2020): 1-8.