IECE

IECE Transactions on Internet of Things
http://dx.doi.org/10.62762/TIOT.2024.128976

RESEARCH ARTICLE

Machine Learning-Based Prediction of Cardiovascular Diseases

Weicheng Sun[1,*], **Ping Zhang**[1,2], **Zilin Wang**[1] **and Dongxu Li**[2]

[1] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2] School of Computer, BaoJi University of Arts and Sciences, Baoji 721016, China

Abstract

With the rapid development of artificial intelligence, extracting latent information from medical data has become increasingly critical. Cardiovascular disease is now a major threat to human health, being one of the leading causes of mortality. Therefore, developing effective prediction methods for cardiovascular diseases is urgently needed. Current medical approaches primarily focus on disease detection rather than prediction, which limits early intervention. By leveraging computational methods, it is possible to predict cardiovascular disease in advance, enabling timely treatment and potentially reducing the disease's impact. This study employs machine learning techniques, including Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), to predict cardiovascular diseases as classification problems. These machine learning models are supported by robust mathematical theory, allowing them to handle non-linear classification challenges effectively. The results offer valuable insights for the prevention and early treatment of cardiovascular diseases.

Keywords: SVM, Random Forest, Machine learning, Cardiovascular disease prediction.

Academic Editor:
Jinchao Chen

Submitted: 19 March 2024
Accepted: 07 May 2024
Published: 23 May 2024

Vol. 2, **No.** 2, 2024.
10.62762/TIOT.2024.128976

***Corresponding author:**
✉ Weicheng Sun
SWCh344@hotmail.com

Citation
Sun, W., Zhang, P., Wang, Z., & Li, D. (2024). Machine Learning-Based Prediction of Cardiovascular Diseases. IECE Transactions on Internet of Things, 2(2), 50–54.

1 Introduction

In the field of medicine, cardiovascular disease has been major causes of death. Due to the improvement of people's living standards and the increase of life pressure, the number of people suffering from cardiovascular diseases is also increasing year by year. Cardiovascular diseases mainly include two types, namely acute cardiovascular diseases and chronic cardiovascular diseases. Traditional wet-lab experiments used for identifying cardiovascular diseases tends to be inefficient and time-consuming, therefore the computer method plays an important role in the treatment of cardiovascular diseases. Chen et al. used cox proportional risk regression model and BP neural network to predict the factors affecting the disease of the elderly, which provided a good reference for the prediction of cardiovascular disease [1]. In addition, Zhang et al. adopted linear regression model to predict the causes of childhood obesity and researched the main causes of childhood obesity from 12 characteristics [2]. At the same time, Li et al. performed logical regression to analyze the factors affecting health of residents to improve the level of oral health and health needs of residents [3]. On the other hand, with the rapid development of machine learning, Aditi et al. used machine learning algorithm to predict heart disease. Firstly, they collected the user's age, sex, blood pressure, heart rate and other data information, and then use multi-layer net to build a machine learning model to predict heart disease. The model has achieved good performance and can be used to predict heart disease [4]. Mrunmayi et

al. constructed a model based on support vector machine to predict diseases. Support Vector Machine is suitable for dealing with high-dimensional data, but it exists the problem of high computational complexity, which makes the model obtain a poor accuracy [5]. Prof. Dhomse Kanchan et al. adopted Random Forest, Support Vector Machine and Bayesian network to predict diabetes. Because of the high dimension of the data, principal component analysis is introduced to reduce the dimension of the data. The results show that the Support Vector Machine has better results [6]. At the same time, Senturk et al. used decision tree, Support Vector Machine, Bayesian network, linear regression and other methods to predict breast cancer. The results indicates that the accuracy of support vector machine is significantly higher than other methods [7]. Satyabrata Aich et al. adopted PCA and decision tree to predict Parkinson's disease, which can help medical staff to detect the disease as soon as possible [8]. Besides, Wen et al. constructed an automatic classification framework for accurate recognition of TS children with multiple types of features. Experimental results show that the model can classify TS children and healthy children with high accuracy [9].

In this paper, we adopted machine learning-based method including SNM, RF and LR to predict cardiovascular disease. We used serval integration features to combine the feature vectors, including summing them up to take the average, taking the geometric mean, choosing the maximum value, and applying the different classifier to train the predictive model. Specifically, we calculate the correlation coefficient for feature extraction, then used different optimize strategies such as maximum likelihood function, cross entropy and sigmoid to obtain better accuracy values, the results indicates that our method get better performance for medical data. The application of machine learning-based method [14, 15] can provide support for prediction of diseases.

2 Materials and methods

The experimental data come from the Svetlana Ulianova research group. The data set mainly includes 12 characteristics, such as age, systolic blood pressure and diastolic blood pressure obtained by sensors. A total of 70,000 pairs of data are collected. Considering the influence of the noise of the data, we perform data pre-processing including the processing of redundancy data and data normalization. Besides, we redefine the label, specifically, when the patient has a disease, we set the label as 1, and otherwise as 0.

2.1 Support vector machine

Support vector machine is based on VC dimension theory and structural risk minimization theory. It can find a best compromise according to the complexity of the limited sample information in the model and learning ability to obtain a better generalization ability [10]. Support Vector Machine is mainly to solve a quadratic programming problem, for quadratic programming problem, there are many methods can be used, but when there are many training samples, the algorithm will face the disaster of dimensionality, which makes it difficult to use support vector machine for classification [11].

Support vector machine first non-linearly maps the training set to a high-dimensional space, so that the linear inseparable data under low-dimensional conditions can be mapped to the high-dimensional space to make it linearly separable. As a result, the hyperplane which can correctly divide the data set and has the largest interval can be solved. The hyperplane can be represented by the following formula.

$$w^T x_i + b = 0 \tag{1}$$

The distance from X to the hyperplane at any point in the sample space can be expressed by the following formula.

$$\gamma = \frac{\left|w^T x + b\right|}{\|w\|} \tag{2}$$

Suppose that the hyperplane can classify the samples correctly. There is the following formula.

$$\begin{cases} w^T X_i + b \geq +1, y_i = +1 \\ w^T X_i + b \leq -1, y_i = -1 \end{cases} \tag{3}$$

The support vector refers to several training samples that are closest to the hyperplane that can make the equal sign valid, so we can find the sum of the distances from the two heterogeneous support vectors to the hyperplane (i.e., the interval). The task of support vector machine is to find the hyperplane with the maximum interval, which can be expressed by the following formula.

$$\max \frac{2}{\|w\|} \left(\text{s.t. } y_i \left(w^T X_i + b \geq 1 \right), i = 1, 2, \ldots, m \right) \tag{4}$$

Therefore, for the support vector machine, after the training is completed, most of the training samples do not need to be retained, and the final model is only related to the support vector.

2.2 Logical regression

Logical regression is a multiple linear regression model. By using this method, we can get weights of independent variables, thus we can obtain the information which relevant factors influence on the experimental results. At the same time, Logical regression usually assumes that the data obeys Bernoulli distribution, then the maximum likelihood function method and the gradient descent method is used to further optimize the model and obtain the optimal parameters [12]. By using logical regression, the training speed of the model is faster, the form is relatively simple, and the explanation is strong. In particular, we use the formula 5 to describe logistical regression

$$f(x\theta) = \frac{1}{1 + e^{-x\theta}} \tag{5}$$

2.3 Random forest

The random forest adopts the ensemble method, which based on decision trees, and the output category of the random forest is determined by the number of individual tree output categories [13]. Meanwhile, with the development of internet of things, advanced WSN-based data transmission technology can transfer and store big data effectively and safely thus satisfy the need of high quality medical data (obtained by medical sensors) fit into the random forest classifier. Specifically, decision trees select the optimal features from current feature sets. However, Random Forest random selects a subset from a node set which belongs to the base decision trees, then chooses optimal features from the subsets. Random forest has high accuracy and can be used in big data set, but also has a good performance for some high-dimensional samples.

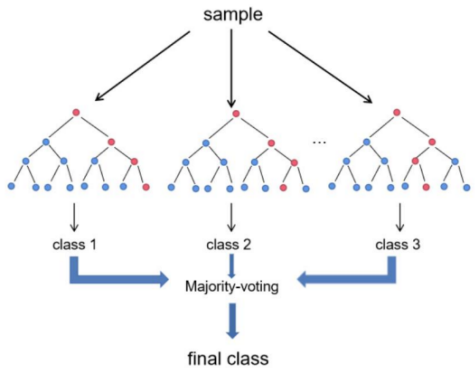

Figure 1. Workflow of RF Classifier

2.4 Prediction assessment

In this study, we used 5-fold cross-validation to train model. Specifically, our data was separated into five subsets, where each subset maintains data distribution consistency. By splitting 5 parts (i.e., four parts for training and one part for testing), our samples can be trained well and then we choose the mean value of 5-fold cross-validation as the final results.

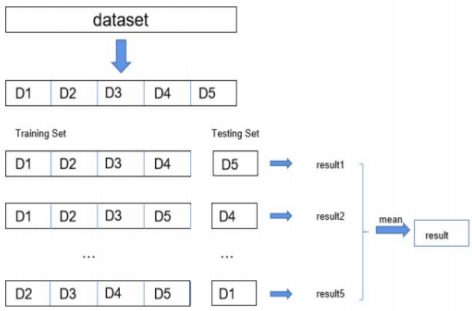

Figure 2. 5-fold cross-validation

In order to evaluate the performance of our model, we calculate accuracy, recall and specificity using following formulas.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

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

$$specificity = \frac{TN}{TN + FP} \tag{8}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives and FN represents the number of false negatives.

3 Experiments

In this experiment, we mainly use correlation coefficient methods for feature extraction. By calculating the correlation between each feature and the disease, we select age, weight, cholesterol, height as the final feature. Before feature extraction, considering the dimension of the data, we first standardize the data so that the data is under the same dimension, mainly using the corresponding package in Sklearn. Here we mainly calculate the correlation of the data, as shown in Table 1 and Table 2.

Table 1. Correlation coefficient(1).

	Age	Gender	Height	Weight
Disease	0.238	0.008	-0.010	0.181

Table 2. Correlation coefficient(2).

	Ap_hi	Ap_lo	Alco	Smoke
Disease	0.054	0.657	-0.035	-0.015

As shown in Figure 3 and 4, we found that there exists obvious correlation between age and disease (table1). In order to further analyze the associations between different factors and disease, we performed descriptive analysis. Here we used boxplot to describe the distribution of age, height and weight, the experimental results are shown in the following figure, we select the appropriate features by setting the threshold as the input of the model. The threshold is set to 0.05. Then the features are selected for subsequent classification.

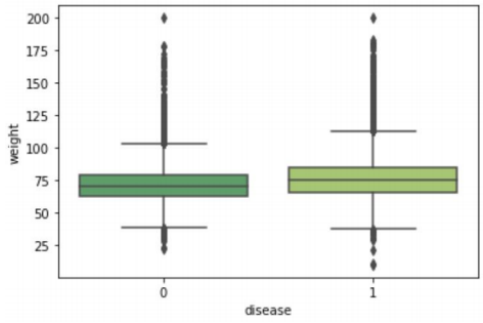

Figure 3. The box plot about weight

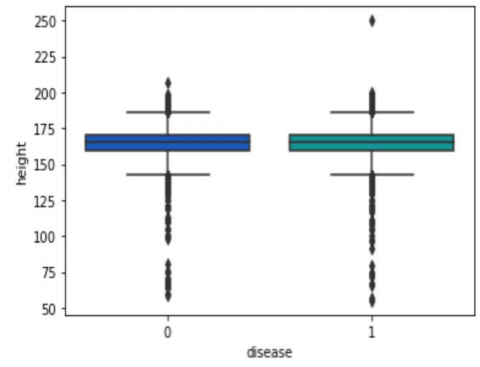

Figure 4. The box plot about height

We use SVM, Random Forest Classifier and Logical Regression Classifier to train the model under 5-fold cross-validation, and the results are shown in following table.

Compared with LR and RF, SVM obtain a better performance. Specifically, SVM, LR and RF achieved the average areas under the ROC curve of 78.84%, 78.41% and 77.50% under 5-fold cross-validation, respectively. On the other hand, we used the following figure to further describe the results of 5-fold cross-validation.

Table 3. 5-fold cross-validation results

Classifier	AUC(%)	Recall(%)	ACC(%)
SVM	78.84	71.47	71.48
LR	78.41	71.62	71.63
RF	77.50	70.67	70.50

4 Conclusion

This paper mainly used the method based on correlation coefficient for feature extraction, then adopted machine learning-based method including SVM, RF and LR to predict cardiovascular disease. The results show that SVM-based method can achieve the average area under the ROC curve of 78.84% under 5-fold cross-validation, which indicates the performance of SVM is better than LR and RF for our data. The use of machine learning-based method can provide support for prediction of diseases.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgement

This work was supported without any funding.

References

[1] Chen, J.h., Study on early warning Model of Ischemic Cardiovascular and Cerebrovascular Diseases in elderly Health Care population. 2010, The third military Medical University.

[2] Zhang, Y.l. and H. Luo, Multiple linear stepwise regression analysis of obesity factors in obese children. *Practical preventive medicine*, 2008. 15(005): p. 1457-1459.

[3] Li, G., Research on Status Evaluation of Oral Health Service and Prediction of Oral Health Human Power. 2004, *Sichuan University*.

[4] Gavhane, A., et al. Prediction of Heart Disease Using Machine Learning. in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2018.

[5] Patil, M., et al. A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine. in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2018.

[6] Kanchan, B.D. and M.M. Kishor. Study of machine learning algorithms for special disease prediction using principal of component analysis. in *International Conference on Global Trends in Signal Processing*. 2017.

[7] Senturk, Z.K. and R. Kara, Breast cancer diagnosis via data mining: performance analysis of seven different

algorithms. *Computer & Engineering*, 2014. 4(1): p. 35-46.

[8] Aich, S., et al. A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data. in *2018 20th International Conference on Advanced Communication Technology* (*ICACT*). 2018.

[9] Wen, H., et al., Multi-modal multiple kernel learning for accurate identification of Tourette syndrome children. *Pattern Recognition*, 2016: p. S0031320316302813.

[10] Suykens, J. Nonlinear modelling and support vector machines. in *IEEE Instrumentation & Measurement Technology Conference*. 2001.

[11] Suykens, J., Support Vector Machines: A Nonlinear Modelling and Control Perspective. *European Journal of Control*, 2001. 7(2–3): p. 311-327.

[12] Wolfe, R.A. and R.L. Strawderman, Logical and statistical fallacies in the use of Cox regression models. *American Journal of Kidney Diseases*, 1996. 27(1): p. 124-129.

[13] Breiman, L., Random forest. *Machine Learning*, 2001. 45: p. 5-32.

[14] Fang, F. A. N. G., Tan, W., & Liu, J. Z. (2005). Tuning of coordinated controllers for boiler-turbine units. *Acta Automatica Sinica*, 31(2), 291-296.

[15] Wang, N., Fang, F., & Feng, M. (2014, May). Multi-objective optimal analysis of comfort and energy management for intelligent buildings. In *The 26th Chinese control and decision conference* (*2014 CCDC*) (pp. 2783-2788). IEEE.

Weicheng Sun Currently studying in the first year of BS in Huazhong Agricultural University. The research direction is bioinformatics, machine learning and graph neural network.

Ping Zhang Currently studying in the first year of PhD in Huazhong Agricultural University. He is a lecturer in Baoji University of Arts and Sciences. His current research interests include bioinformatics, machine learning and graph neural network.

Zilin Wang Currently studying in the first year of BS in Huazhong Agricultural University. The research direction is bioinformatics

Dongxu Li Graduated with B.S at the Department of Computer of Baoji University of Arts and Sciences from 2018 to 2022. His current research interests include machine learning and computer vision.