Journal of Medical Systems (2019) 43: 228
https://doi.org/10.1007/s10916-019-1346-x

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Computer-Aided Diagnosis and Clinical Trials of Cardiovascular Diseases Based on Artificial Intelligence Technologies for Risk-Early Warning Model

Bin Li[1,2] · Shuai Ding[2] · Guolei Song[1] · Jiajia Li[1] · Qian Zhang[1]

Received: 17 February 2019 / Accepted: 20 May 2019 / Published online: 13 June 2019

Abstract
The use of artificial intelligence in medicine is currently an issue of great interest, especially with regard to the diagnostic or predictive analysis of medical data. In order to achieve the regional medical and public health data analysis through artificial intelligence technologies, spark data analysis is adopted as the research platform for hypertension patients, and artificial intelligence technologies are used to preprocess the data with inconsistency, redundancy, incompleteness, noise and error; Aiming at the unbalanced data sets, the Z-score standard is adopted to convert data into usable form suitable for data mining. And, the application of Logistic, Naive Bayesian regression, and support vector machine based on three groups of different prognosis in severe cases, including stroke, heart failure and renal failure symptoms, establish the risk early warning model for 3 years time. In addition, to select the optimal feature subset based on medicine big-data features, the model simplification and optimization are done in training process, the experimental results show that the feature subset selection can ensure the classification performance similar to the clinical features of the model. Therefore, according to chronic cardiovascular disease, acute cardiovascular events and cardiovascular events caused by critical illness events, we screen out the relevant prognosis of serious illness (stroke, heart failure, renal failure), which is related to the prognosis of serious illness. Targeted prevention has a guiding role and practical significance according to the results of artificial intelligence analysis.

Keywords Chronic cardiovascular disease · Artificial Intelligence · Z-score standard · Logistic · Naive Bayesian regression · Support vector machine · Clinical feature

Introduction

With the development of information technology in medical industry, a large amount of clinical medical data has been generated and continued to explode every year. Massive disease diagnosis data, clinical trial data and resident behavioral health data are together to open up the medical reform for the era of big data. Artificial intelligence technology can predict development trend and potential rules of medical data. This has a very important value for disease diagnosis and treatment and medical research [1].

Cardiovascular disease is a major public health problem that threatens global public health and affects seriously the healthy development of human health, accounting for about 30,070 of the world's deaths in worldwide [2]. Hypertension is a common chronic cardiovascular disease, leading easily to stroke, heart failure, renal failure and the prognosis of severe disease, and these symptoms are associated with a variety of the risk factor [3, 4]. Early discovery, effective prevention, control of these risk factors and the establishment of a serious prognosis of the risk early warning model have important significance to control the onset of cardiovascular disease.The researches of vascular disease at home and abroad, are mostly based on the set of experimental programs

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Bin Li
libin2010000@163.com

[1] The First Affiliated Hospital of Bengbu Medical College, Bengbu 233004, Anhui, China

[2] School of Management HeFei University of Technology, Hefei 230009, Anhui, China

and selecting specify populations and the possible risk factors, and can not fully complete effectively grasp the pathogenic factors. Most of the methods based on medical statistics has some limitations [5].

Recently, domestic and foreign scholars mainly use existing artificial intelligence algorithms for medical data analysis [6, 7]. Since artificial neural network is the forefront interdisciplinary subject with rapid development of artificial intelligence in the world, it has strong learning and computing power, and can better adapt to the change of data space [8]. Its application opens up a new way for the research of artificial intelligence. Therefore, Hudson et.al establish diagnosis model of diabetes based on BP neural network [9], where the experimental data are extracted and integrated, and input feature vectors are selected to build neural network models. In order to improve learning efficiency, the parameters of the model were adjusted [10]. The feasibility of applying BP neural network to diabetes diagnosis is verified through experiments [11].

In order to improve the accuracy and speed of diagnosis, Literature [12] proposed a diagnosis model of diabetes based on probabilistic neural network(PNN). The PNN with simple network structure and concise training is selected to establish diabetes diagnosis model. After comparison of the two models, it is found that PNN needs less parameters to adjust, and does not need to determine the network structure of the hidden layer and the number of hidden layer neurons, which is easier to be implemented and used than the BP network. In addition, the average accuracy and the corresponding standard deviation of the 20 tests are also calculated, it shows that PNN model is more suitable for building diabetes diagnosis model. [13] Compared with the traditional diagnosis process, it can effectively save the doctor's time and improve the diagnosis efficiency. However, these algorithms are greatly interfered by noise, and the wrong samples directly affect the classification results.

In order to improve the diagnosis efficiency in imbalance and unstructured data, an intelligent medical service system based on classification algorithm is proposed in [14]. Based on the research of medical diagnosis expert systems at home and abroad, the principle of artificial intelligence medical diagnosis is summarized: the symptoms of patients are classified as known diseases. Based on the classification analysis technology in data mining, the inference engine is realized to classify diseases. In order to obtain better diagnostic results, association analysis is used to mine various potential factors behind the disease, and a combinational classification algorithm based on association rule mining to obtain latent factorswas proposed to improve the optimization [15]. The algorithm solves the problem of overfitting of the single decision tree [16]. In addition, the strong influence of individual differences on the diagnostic classification is used to optimize the voter's voting results for the combined classifier results through the latent factor. Finally, the classification result with the highest confidence level is obtained. And experiments were conducted using real medical big data, which proved that the inference engine based on classification algorithm can meet the needs of intelligent medical diagnosis. Then, based on the Java Web technology stack, the entire smart medical service system was implemented, including a front-end service system that provides patients with functions such as intelligent diagnosis, reservation, inquiry, and message system, and a back-end management system that provides management services for the unmanaged administrator [16]. Moreover, privacy data is protected based on digital watermarks. The functional and performance tests of the entire system were both passed [16, 17].

As can be seen from the above analysis, computer-aided diagnosis and clinical trials of cardiovascular diseases based on artificial intelligence technologies is divided into two parts [18]. The first part includes the distributed storage, data cleaning, data mining technology, which is to achieve the effective integration of scattered data and can extract medicine data by making the software model of deep learning history of medicine data [14]. The second part is to compare the calculation of the model and to describe in detail the remote cardiovascular medicine data [19].

This paper is for the regional medical and public health data analysis and demonstration project application research through the data mining technology, using spark data analysis as the research platform for hypertension patients [20]. The main research contents and contributions of this paper include the following aspects Based on the Spark data processing platform [21], data mining technology is used to preprocess the data with inconsistency, redundancy, incompleteness, noise and error; Since medical data has unbalancedness [22], namely the number of the patient population sample is far less than the number of samples of healthy people. Therefore, the stratified sampling technique is used to transform the unbalanced data sets into balance data, and the Z-score standard is also used to convert data into usable form suitable for data mining. The application of data mining algorithm based on three groups of different prognosis in severe cases, including stroke, heart failure and renal failure symptoms, establish the risk early warning model for 3 years time. The experimental results show that compared with the traditional risk model, our proposed model has practical significance with better prediction effect. In addition, The application of chi-square test, with p value<0.05 as the clinical reference standard, is to select attributes with clinical significance. To select the optimal feature subset selection algorithm based on SVM-RFE features [23], model simplification and optimization are introduced, where the experimental results show that the decrease in the feature space dimension significant case, feature subset selection to ensure the classification performance similar to the the clinical features of the model. Therefore, according to chronic cardiovascular disease and acute cardiovascular events and cardiovascular events caused by critical illness events, we

screen out the relevant prognosis of serious illness (stroke, heart failure, renal failure) for the main risk factor, which is related to the prognosis of serious illness. Targeted prevention has a guiding role and practical significance based on the results of artificial intelligence analysis.

Artificial intelligence technologies

Logistic regression

Logistic is the most classical classification method in machine learning [24]. In essence, it uses Sigmoid function normalized linear regression to reduce the prediction range and limit the prediction value to [0,1] interval as classification model. The function F(x) is a Logistic function and f(x) is a Density function, where F(x) is a S-shaped curve centered on a point (0,5). Logistic formula is written as follows:

$$\mathrm{F}(x) = P(X \leq x) = 1/(1 + \exp(-(x-u)/\gamma)) \tag{1}$$

Therefore, the conditional probability distribution of binomial logistic regression model can be expressed as:

$$\mathrm{P}(Y = 1|x) = \frac{\exp(w \otimes x + b)}{1 + \exp(w \otimes x + b)} \tag{2}$$

where $x \in R^n$ is input vector; $Y \in \{0, 1\}$ is output target value; $w \in R^n$ and $b \in R$ are parameters; w denotes weight vector; b represents offset; $w \otimes x$ is the inner product of w and x. The equation form of $P(Y = 0|x)$ is similar.

Give a training set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, where $x_i \in R^n$, $y_i \in \{0, 1\}$. According to Logistic regression model, the Logarithmic likelihood function is obtained, which is denoted as

$$\begin{aligned} \mathrm{L(w)} &= \sum_{i=1}^{N} \left(y_i \log(P(Y = 1|x_i) + (1-y_i)\log(1-P(Y = 1|x_i))) \right) \\ &= \sum_{i=1}^{N} (y_i(w \otimes x_i) - \log(1 + \exp(w \otimes x_i))) \end{aligned} \tag{3}$$

The maximum value of *L(w)* can be found, namely, the estimated value of W. Thus the problem becomes an optimization problem for solving logarithmic likelihood function as objective function.

Naive Bayes algorithm

Naive Bayes Algorithm is a simple classification method based on the assumption that Bayesian principle and feature conditions are relatively independent [25]. Given a training set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, it is assumed that the joint probability distribution of learning output and input by taking advantage of features which are relatively independent. Then the posterior maximum probability is obtained by Bayesian theorem according to the new sample X of input.

Suppose input sample set $x \in R^n$, the output class labeled set $S = \{s_1, s_2, \cdots, s_k\}$, and given that X is a random variable on input sample set χ, Y is a random variable on output class labeled set B, and $P(x, y)$ is defined as the joint distribution of x and y. Thus Naive Bayesian classifier can be expressed as:

$$y = f(x) = arg\max_{c_k} \frac{P(Y = c_k) \prod_j P\left(X^{(j)} = Y^{(j)}|Y = c_k\right)}{\sum_k P(Y = c_k) \prod_j P\left(X^{(j)} = Y^{(j)}|Y = c_k\right)} \tag{4}$$

It is noted that the denominator in the upper formula is the same for all class labels, so we can obtain

$$y = arg\max_{c_k} P(Y = c_k) \prod_j P\left(X^{(j)} = Y^{(j)}|Y = c_k\right) \tag{5}$$

Support vector machine

The basic model of the support vector machines (SVM) algorithm is a linear classifier that defines the maximum spacing in the feature space [26]. The kernel technique is introduced to solve the nonlinear problem, making it a substantially nonlinear classifier. The learning strategy of support vector machine is to find the separated hyperplane that maximizes the interval between different types of samples, and then abstracts it into the solution optimization problem, which is equivalent to the regularization of the hinge loss function minimization problem。 .

This automatic classification experiment used LIBSVM [27], which is a software package of SVM pattern recognition and regression developed and designed by Lin Chih-Jen of the University of Taiwan. In the Windows series system, the execution files can be directly called and the problems, such as c-SVM, logic-SVM, SVR and logic-SVR, can be solved. The software has relatively fewer parameter adjustments involved in SVM and provides many default parameters, which can be used to solve many problems. Currently, there are two popular strategies of SVM multi-classification: one-with-all and one-with-one. The multi-classification in LIBSVM adopts a one-with-one strategy [28]. There were 136 data samples used in the study, which were provided by the surgeons. The LIBSVM toolkit was adopted for training, and the total number of support vectors was 136. In this paper, the accuracy rate was used as the evaluation index of classification.

Feature selection

Chi-square test is an important feature selection algorithm, which is a hypothesis test in statistics. Chi-square statistics χ^2 is used to measure the correlation between attributes and classes. Chi-square test is usually used to calculate *P* Value by statistical analysis of medical experimental data [29]. Chi-

square test assumes that events A and B are relatively independent. In the case of binary classification, events A and B can be considered as features and classes. Firstly, we assume that there is no correlation between field features A and class B. Then we calculate the chi-square statistics between each field feature and class. The chi-square formula is as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_o - f_c)}{f_c} \tag{6}$$

where f_o represents the actual number of occurrences, f_c represents the expected number of occurrences by hypothesis theory. Chi-square statistics indicate the degree of deviation between the observed value N and the expected value E [30]. If the value is larger, it means that the independent hypothesis between features and classes is not valid, that is, there is a strong correlation between features and classes. Then the chi-square statistics of classes and features are calculated, and then the P Value are calculated by degrees of freedom. Finally, according to the order of P Value from small to large, screen the features below the given threshold. The threshold is normally set to 0.05.

Medical data processing

Data processing is the most important part of data mining. According to the characteristics of medical big data, this paper uses Spark platform for interactive query and data analysis, Oracle database as an assistant exploration and data validation. In this paper, data processing is mainly aimed at critical diseases caused by cardiovascular disease and cardiovascular function damage. Guided by medical knowledge, some field attributes which are not related to mining objectives are discarded, so as to provide more effective data for data mining, reduce the amount of data processing, improve the accuracy of knowledge discovery and the effectiveness of data mining.

Data transformation

Data transformation is to transform data into usable forms suitable for data mining models. Data transformation mainly includes the following contents:

1) Smooth: remove noise from data, including regression and clustering.
2) Aggregation: collection or aggregation of data.
3) Data generalization: using conceptual hierarchy to replace underlying or "raw" data with high-level concepts.
4) Normalization: it scales attribute data in proportion to a specific interval.
5) Attribute construction: construct new attributes to be added to the attribute set to help the mining process.

In this paper, data transformation of medical data mainly involves two situations: data normalization [31] and attribute construction. Data normalization can be reduced to a small interval, such as 0-1 interval, by scaling the attribute values in proportion.

In addition, z-score normalization is used to normalize the value of attribute A by calculating the mean and standard deviation of attribute A, which is expressed by the following formula:

$$\tilde{v} = (v - u)/\sigma_A \tag{7}$$

where u and σ_A represent the mean and standard deviation of attribute A, respectively; v represents the value of attribute A; $\tilde{v}$ represents the value of normalized A. Another attribute construction is to construct and add new attributes by given attributes. In this paper, there are text data fields, which need to be converted into numerical data to meet the requirements of the model.

Data integration

Data integration mainly integrates data from multiple data sources and integrates them into a unified data set physically or logically [30]. There are many problems to be considered in data integration. The first is pattern integration and object matching, that is to say, the matching and recognition of equivalent entities from multiple data sources. This requires judging whether the attributes of different data sources are the same according to the meta-information of each attribute. Another important problem is redundancy. Detecting redundancy, based on available data, mainly measures how much one attribute can contain another attribute. For numerical attributes, the correlation $r_{A,B}$ between attributes A and B is estimated by calculating the correlation coefficients between the two attributes, i.e.:

$$r_{A,B} = \frac{\sum_{i=1}^{N} \left(a_i - \overline{A}\right)\left(b_i - \overline{B}\right)}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N} (a_i b_i) - \overline{A}\,\overline{B}}{N\sigma_A\sigma_B} \tag{8}$$

where N is the number of samples; a_i and b_i respectively are the value of A and B in sample i; $\overline{A}$ and $\overline{B}$ are the mean of A and B, respectively; σ_A and σ_B are the standard deviation of A and B, respectively. If $r_{A,B}$ is greater than 0, A and B are correlated. The greater the value is, the stronger the correlation is. Therefore, higher $r_{A,B}$ value indicates that A or B can be deleted as redundancy. For discrete data, the continuity between A and B is found by chi-square test. The third important issue is the detection and processing of data value conflicts. For example, in the real world, the attribute values of different data sources may be different for the same entity, which requires manual identification and judgment processing.

Data reduction

Data sets generally contain a large number of features, and the number of samples is also very large. Data reduction is to maximally simplify the amount of data while keeping the original data as much as possible. Common data reduction methods are:

1) Data cube aggregation: the method of aggregation is applied to data cube, which is the multi-dimensional modeling and representation of data.
2) Dimensional protocol: detecting and deleting irrelevant, weak correlation or redundant attributes. Common methods include principal component analysis, etc.
3) Numerical protocol: data are represented by smaller data. Common methods include histogram, clustering, sampling, etc.
4) Discrete and conceptual hierarchical generation: by dividing attribute value fields into intervals, the number of given continuous attribute values is reduced, and the label of intervals replaces the actual data values.

The data protocol in this paper mainly involves two aspects: dimension and numerical reduction, which are aimed at the records and characteristics of the original data set. Dimension reduction mainly uses chi-square test and SVM-RFE for feature selection [32]. The numerical reduction mainly uses sampling technology, because it enables much smaller subsets of samples to represent large data sets. There are four most commonly used sampling methods, including simple random sampling without playback, simple random sampling with playback, cluster sampling and stratified sampling. In this paper, the experimental group and the control group data sample size is far from each other. In order to avoid data imbalance, stratified sampling of control group data is conducted. According to the principle of stratified sampling and the proportion of age and sex distribution in the experimental group,the same number of samples in the control group are took as that in the experimental group.

Experimental results and analysis

Based on the two types of events, chronic and acute cardiovascular events and critical illness events caused by cardiovascular function loss, a risk early warning model for 3-year long-term prognostic severe disease events was established. In order to make better risk prediction, this paper mainly uses Spark Mllib Library, Logistic Regression, Naive Bayesian and Support Vector Machine to build risk early warning models for the above two types of events. Through parameter selection, the model compares and chooses the best classification early warning model. Finally, based on the Logistic regression model, the main risk factors are selected according to the characteristic weight.

The description of data set

Aiming at two kinds of events, chronic and acute cardiovascular events and critical illness events caused by cardiovascular function loss, this paper takes hypertension patients as the research population. Through data mining, three sets of data sets are extracted from massive medical diagnostic records, including stroke, heart failure and renal failure. Among them, stroke and heart failure belong to acute cardiovascular events, while renal failure belongs to critical illness events caused by cardiovascular function loss. In biomedical research, P Value <0.05 is used as the criterion of clinical significance. Chi-square test is used to screen out the characteristic attributes that tally with clinical significance. The specific data sets are shown in Table 1.

Evaluation index

Classification algorithm is selected for model training. Model training divides the data into training data and test data according to the learning need. The training data is used to learn the model, and the test data is used to predict the validity of the model and to measure the advantages and disadvantages of the model. Aiming at the evaluation index of classification model, this paper mainly adopts accuracy, precision, specificity, sensitivity, AUC, F-Measure.

AUC refers to the area under the ROC curve [33]. ROC curve, also known as receiver operating characteristic curve, first appeared in signal detection and was used to select the optimal signal detection model. Since ROC analysis is not affected by cost/benefit, it

Table 1 Data set description

Severe prognostic illness	Number of samples in EG	Number of samples in CG	Full characteristic number	Clinical characteristic number
Stroke	5257	5296	189	130
Heart failure	714	718	166	94
Renal failure	796	808	159	78

Table 2 Classification results on *stroke* datasets

Algorithms	AUC	Accuracy	Precision	Sensitivity	Specificity	F-Measure
Naive Bayes	0.8407	0.7662	0.8199	0.6811	0.8500	0.7440
SVM(linear)	0.8634	0.7827	0.8265	0.7104	0.8538	0.7641
Logistic Regression	0.8605	0.7820	0.8165	0.7225	0.8406	0.7665

can give objective and neutral evaluation, and has been applied in data mining and machine learning. Where, the ordinate of ROC curve is the true positive rate (TPR) and abscissa is false positive rate (FPR) of classification algorithm. The formula is as follows:

$$\mathrm{TPR} = TP/(TP + FN) \tag{9}$$

$$FPR = FP/(FP + FN) \tag{10}$$

If the target decision-making value obtained by the classification model is to judge the positive or negative class probability of samples, then we can select a probability threshold, which is formed by connecting a series of points under different thresholds. Calculate the area under ROC curve to obtain the AUC value. There are two extreme phenomena in ROC curve: one is that all samples are classified correctly, AUC = 1, the other is completely ineffective, equivalent to random classification, AUC = 0.5, but the normal situation is that ROC curve is between the two. The closer the AUC value is to 1, the higher the classification accuracy is, and the better the model is. When the samples are very unbalanced, the accuracy of classification model will be seriously affected by the distribution of samples. For example, the number of negative samples is 9 times that of positive samples. When the models are predicted to be negative, the accuracy can reach more than 90%, but this is obviously not what we expected. AUC value is not affected by data imbalance and reflects objectivity. It is also the main evaluation index in this paper.

Establishment of early-risk warning model

This paper mainly uses Logistic Regression, Naive Bayesian and Support Vector Machine classification model algorithm to respectively build risk model for chronic and acute cardiovascular events and critical illness events caused by cardiovascular function loss. There are three sets of data sets in total. Spark MLlib Library is used to carry out the above algorithm experiments. For each data set, it is divided into training set and independent test set. The proportion of training data is 70%, and the proportion of test data is 30%. The parameters of the training set are adjusted by ten times cross validation, and the optimal parameters are selected. Then the model is evaluated by independent test set. In the Spark MLlib Library, this paper mainly uses Logistic Regression with LBFS class to build logistic regression model, Naive Bays class to build Bayesian model and SVM with SGD class to build support vector machine model, and uses train function to train model and predict function to test model.

In this paper, we mainly study the diagnostic algorithms of chronic and acute cardiovascular events. We screened out two prognostic severe diseases related to the event, stroke and heart failure. Stroke and heart failure datasets were named respectively according to the final clinical characteristics data. The data set is divided into training set and independent test set according to 7:3 using the retention method. The training set is used to find the optimal parameters and determine the final model. Thirty randomly repeated experiments were conducted to evaluate the model. And the average value was obtained as the result of the model evaluation. This paper uses Logistic regression, Naive Bayesian and Support Vector Machine to establish risk early warning models. Table 2 lists the classification model results of stroke data set. The results show that the AUC value of SVM and Logistic regression reaches 0.86, and Accuracy reaches more than 0.78, which are higher than Naive Bayesian classification model. And Logistic regression is superior to other two classification models in sensitivity. Table 3 lists the classification model results of heart failure dataset. The results show that the AUC values of Naive Bayesian, Logistic regression and SVM are all above 0.92, and the Accuracy values are above 0.85. The three models are relatively stable and have little change. And the value of SVM in sensitivity is obviously superior to the other two classification models. When other indicators are relatively stable, the medical significance of Sensitivity assessment indicators is

Table 3 Classification results on *heart failure* datasets

Algorithms	AUC	Accuracy	Precision	Sensitivity	Specificity	F-Measure
Naive Bayes	0. 9220	0. 8571	0. 8884	0. 8195	0. 8951	0. 8522
SVM(linear)	0. 9246	0. 8595	0. 8679	0. 8501	0. 8695	0. 8584
Logistic Regression	0. 9269	0. 8529	0. 8915	0. 8048	0. 9014	0. 8456

important, indicating the percentage of actual diseases correctly diagnosed. Figures 1 and 2 show the ROC curves of *stroke* and *heart failure* datasets on different models intuitively.

In this paper, SVM-RFE feature selection algorithm is used to select the optimal subset of stroke and heart failure datasets, and 10, 20 and 30 feature subsets are selected from the full features of datasets. The performance of different feature subsets under three models is verified, and the AUC and Sensitivity indexes of each model are compared. Table 4 and Fig. 3 show the results of *stroke* dataset based on SVM-RFE feature selection algorithm in different models. It can be seen that the AUC value of Logistic regression can reach 0.85 when selecting 70 feature subsets, which is close to the performance of *Stroke* with full clinical feature model. It shows that the feature subset selected by SVM-RFE algorithm can fully represent the original full clinical feature data. Table 4 and Fig. 3 show the results of heart failure dataset based on SVM-RFE feature selection algorithm in different models. It can be seen that when Logistic regression and SVM select only 50 feature subsets, the AUC value can reach above 0.91, which is very close to the performance of heart failure with full clinical feature model; and the Sensitivity value is the best in SVM model, reaching 0.8625, which is higher than 1 percentage point in full feature model. It shows that the feature subset selected by SVM-RFE algorithm can fully represent the original clinical feature data.

Obviously, we use big data mining method to analyze the clinical data comprehensively. Compared with the traditional medical experimental method, the risk-early warning model of severe prognosis of stroke and heart failure is more effective, higher than the evaluation indicators of relevant studies at this stage. For example, the AUC value of logistic regression and SVM models is above 0.86, which is higher than that of stroke model proposed by Stevens [34] et al. with the highest AUC value of 0.769, higher than Aditya Khosla's five-year risk early warning model with the highest AUC value of 0.777 and Sensitivity of 0.71, much higher than literature [35] which is one-year stroke risk model for 0.41. Similarly, for the severe prognosis of heart failure, the AUC value of the model is above 0.92 and Accuracy value is above 0.85, while the accuracy of Kenney Ng's early warning risk model for heart failure is only 0.83; the AUC value of Fen Miao [36]'s one-year short-term risk early warning model for heart failure is only 0.712.

Major risk factors for heart failure

Heart failure refers to cardiac insufficiency. When proper venous reflux occurs, the cardiac output cannot meet the needs of normal tissue metabolism due to dysfunction of cardiac diastolic or systolic function. There are many causes of heart failure, including myocardial diastolic dysfunction, myocarditis, heart disease,

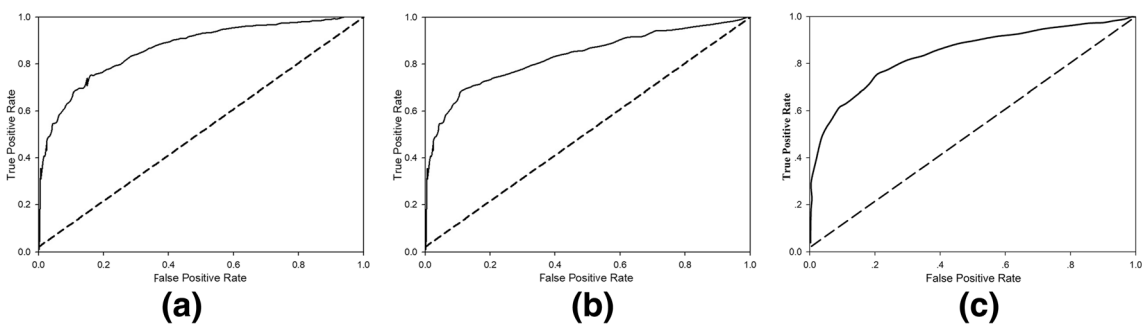

Fig. 1 ROC curves for stroke datasets; **a** Logistic Regression; **b** Naive Bayes; **c** SVM(linear)

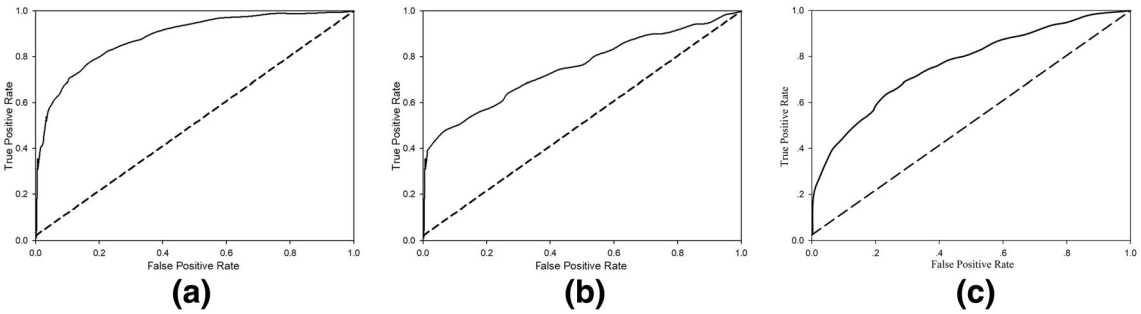

Fig. 2 ROC curves for *heart failure* datasets; **a** Logistic Regression; **b** Naive Bayes; **c** SVM(linear)

metabolic dysfunction, pulmonary artery stenosis. Table 5 shows that the sum of weighted squares of the first nine features contributes about 80% and can be used as the main risk factor of heart failure risk early warning model. Table 5 shows the characteristic weights of nine risk factors and the specific conditions of disease types.

Positive correlation indicates that the risk of heart failure increases with the onset of disease associated with risk factors. The ICD codes studied in this paper only correspond to 0 and 1. If a patient with hypertension suffers from the disease type corresponding to all the risk factors mentioned above, it indicates that the patient is likely to be in the stage of heart failure or develop into heart failure in the future.

Clinical trials

We mainly analyzed the distribution of the first three risk factors of chronic cardiovascular disease by single variable analysis. It is shown in clinical trials that the number of patients with cardiovascular disease in the experimental group is higher than that in the control group to a certain extent, and the number of patients without disease in the control group is higher than that in the experimental group to a certain extent, which conforms to the results of the model. It shows that cardiovascular disease increases the possibility of heart failure to some extent. Clinical experiments show that myocardial diastolic dysfunction is a cardiovascular pathogenic factor, and cardiovascular disease is likely to deteriorate into heart failure. Therefore, early diagnosis and prevention are of great significance to the prevention and control of heart failure. And the results of clinical comparative validation show that the use of artificial intelligence algorithm is helpful to mine disease information and improve the accuracy of diagnosis.

Conclusions

This paper establishes risk early warning models for three groups of severe prognostic diseases based on chronic and acute cardiovascular events, and critical illness events caused by cardiovascular function loss. In

Table 4 AUC and Sensitivity results based on SVM-RFE feature selection on *Stroke* dataset

SVM-RFF	Naive Bayes		SVM(linear)		Logistic Regression	
	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
10	0.5418	0.2024	0.6344	0.3091	0.6358	0.3153
20	0.5739	0.4065	0.7937	0.7058	0.7505	0.5450
30	0.5752	0.5737	0.8341	0.7199	0.8061	0.5928
40	0.5755	0.5618	0.8119	0.7963	0.8246	0.6421
50	0.7156	0.5748	0.8290	0.7155	0.8330	0.6665
60	0.7259	0.5855	0.8406	0.7226	0.8509	0.7129
70	0.7986	0.6657	0.8511	0.6976	0.8579	0.7156
80	0.7843	0.8633	0.7201	0.8552	0.7170	0.8075
90	0.6662	0.8503	0.7134	0.8568	0.7200	0.8025
100	0.8025	0.6560	0.8552	0.7211	0.8596	0.7056

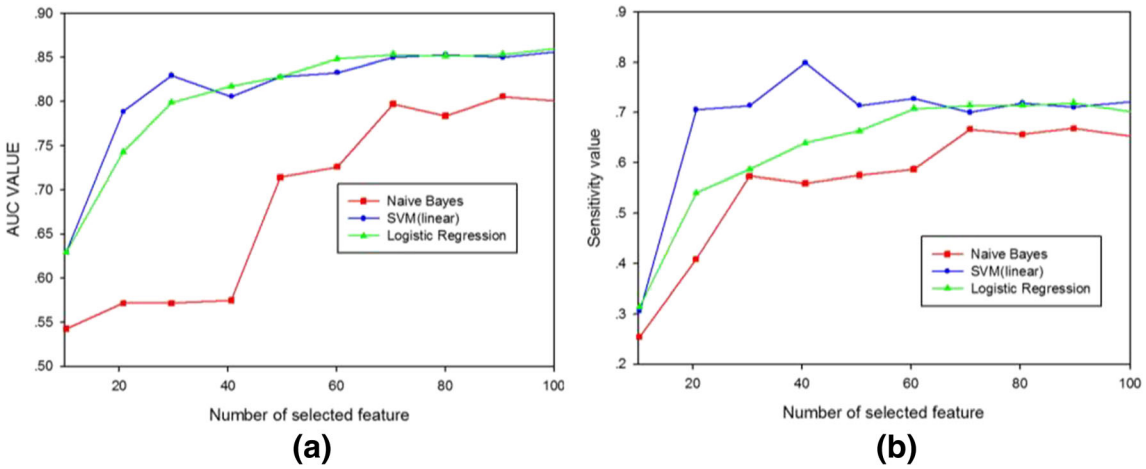

Fig. 3 AUC and Sensitivity curves on *Stroke* dataset. **a** AUC curve; **b** Sensitivity curve

our cardiovascular disease research, there are many serious prognostic diseases, such as acute myocardial infarction, pulmonary embolism, severe brain neurological diseases, etc. We can carry out risk early warning model for a specific complication and explore its risk factors. Secondly, only two feature selection algorithms are used in this study. Chi-square test is used to select features that are clinically meaningful, and the other is used to reduce the dimension to find the optimal feature subset. Although SVM-RFE has a good feature selection effect, it does not compare multiple feature selection algorithms to select the most suitable feature subset for data. So the next research can select the most suitable feature subset by comparing and analyzing different feature selection algorithms. Furthermore, because of the fact that the subjects of this study are hypertensive patients, the number of laboratory data is relatively small, where the laboratory data include many biochemical indicators, such as human nitrite content, glucose content, urinary protein content, etc. Follow-up studies on other diseases, such as diabetics, can add biochemical indicators to model analysis with a certain amount of samples.

Acknowledgements This study was funded by 2017 Social Science Key Project of Bengbu Medical College. (Fund no.: BYKY17146skZD).

Compliance with ethical standards

Conflict of interest We declare that we have no conflict of interest.

Human and animal rights The paper does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Yu-Guang, Y., De-Chang, L. I., Hong-Yu, G. et al., Application of the Artificial Intelligence Technology in Coronary Heart Disease Diagnosis. Changchun: Journal of Changchun Normal University, 2008.
2. Park, S.H., and Han, K., Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 286:800–809, 2018.
3. Nagano, H., Big data, information and communication technology, artificial intelligence, Internet of things: How important are they for

Table 5 Contribution for different number of features

Risk factors	ICD code	Weight	Correlation
i25.103	Coronary atherosclerotic heart disease	1.4088452	Positive
Bit38	Pulmonary edema	0.8934645	Positive
Bit16	Disorder of amino acid metabolism	0.8018865	Positive
Bit33	Viral pneumonia	0.6868133	Positive
I48.×01	Auricular fibrillation	0.6216687	Positive
N18	Chronic kidney disease	0.6205462	Positive
R51	Headache	0.6073420	Positive
I11.900	Hypertensive heart disease	0.5078326	Positive
I42	Cardiomyopathy	0.4853272	Positive

gastroenterological surgery? Annals of Gastroenterological Surgery 2(3):166–166, 2018.
4. Ríos, S.A., Tenorio, F.G., and Jimenezmolina, A., A benchmark on artificial intelligence techniques for automatic chronic respiratory diseases risk classification. In: Kes-inmed-15 Third International Conference on Innovation in Medicine & Healthcare. Cham: Springer, 2016.
5. Drotár, P., Mekyska, J., Rektorová, I. et al., Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif. Intell. Med. 67:39–46, 2016.
6. Canadas, J., Sánchez-Molina, J. A., Rodríguez, F. et al., Improving automatic climate control with decision support techniques to minimize disease effects in greenhouse tomatoes. Information Processing in Agriculture 4(1):50–63, 2017.
7. Yifeng, X. U., Lijun, L., Qingsong, H. et al., Research on TF-IDF weight improvement algorithm in intelligent guidance system. Computer Engineering & Applications, 2017.
8. Dyster, T.G., Sheth, S.A., Mckhann, G.M., et al., Ready or not, here we go: Decision-making strategies from artificial intelligence based on deep neural networks. *Neurosurgery*. 78(6):N11-2, 2016.
9. Hudson, D. L., and Cohen, M. E., Use of intelligent agents in the diagnosis of cardiac disorders. Comput. Cardiol. IEEE, 2002.
10. Ahmad, F., Isa, N. A. M., Hussain, Z. et al., Intelligent Medical Disease Diagnosis Using Improved Hybrid Genetic Algorithm - Multilayer Perceptron Network. J. Med. Syst. 37(2):9934, 2013.
11. Yan, J., Lu, Y., Xu, Y. et al., INTELLIGENT DIAGNOSIS OF CARDIOVASCULAR DISEASES UTILIZING ECG SIGNALS. International Journal of Information Acquisition 07(02):81–97, 2010.
12. Sekar, B. D., and Dong, M., Function Formula Oriented Construction of Bayesian Inference Nets for Diagnosis of Cardiovascular Disease. Biomed. Res. Int. 2014(1):376378, 2014.
13. Sun, B., Li, Y., and Zhang, L., The Intelligent System of Cardiovascular Disease Diagnosis Based on Extension Data Mining. Cutting-Edge Research Topics on Multiple Criteria Decision Making. Berlin Heidelberg: Springer, 2009.
14. Salah, R. B., and Chabchoub, S., Intelligent diagnosis method of cardiovascular anomalies using medical signal processing. World Congress on Information Technology & Computer Applications. IEEE, 2016.
15. Ghareh Baghi, A., and Lindén, M., An Internet-Based Tool for Pediatric Cardiac Disease Diagnosis using Intelligent Phonocardiography. International Internet of Things Summit. New York: Springer International Publishing, 2015.
16. Glass, T. F., Knapp, J., Amburn, P. et al., Use of artificial intelligence to identify cardiovascular compromise in a model of hemorrhagic shock. Crit. Care Med. 32(2):450–456, 2004.
17. Lee, H. G., Noh, K., Lee, B. J. et al., Cardiovascular Disease Diagnosis Method by Emerging Patterns. Advanced Data Mining and Applications. Berlin Heidelberg: Springer, 2006.
18. Filimon, D. M., and Albu, A., Skin diseases diagnosis using artificial neural networks. IEEE International Symposium on Applied Computational Intelligence & Informatics., IEEE, 2014.
19. Valavanis I K, Mougiakakou S G, Grimaldi K A, et al. Analysis of Postprandial Lipemia as a Cardiovascular Disease Risk Factor using Genetic and Clinical Information: An Artificial Neural Network Perspective. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2008:4609-4612, 2008.
20. Feshki, M. G., and Shijani, O. S., Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network. Artificial Intelligence & Robotics. IEEE, 2016.
21. Nes, B. M., Gutvik, C. R., Lavie, C. J. et al., Personalized Activity Intelligence (PAI) for Prevention of Cardiovascular Disease and Promotion of Physical Activity. Am. J. Med. 130(3):328–336, 2017.
22. Tang, Z., Wang, S., Huo, J. et al., Bayesian Framework with Non-local and Low-rank Constraint for Image Reconstruction. J. Phys. Conf. Ser., 2017.
23. Alhadidi, T., and Salah, R. B., A new intelligent method for the automatic diagnosis of cardiovascular anomalies. 2015 17th International Conference on E-health Networking, Application & Services (HealthCom). IEEE, 2015.
24. Bondy, C. A., Congenital Cardiovascular Disease in Turner Syndrome. Congenit. Heart Dis. 3(1):2–15, 2008.
25. Babič, F., Olejár, J., Vantová, Z. et al., Predictive and Descriptive Analysis for Heart Disease Diagnosis. Comput. Sci. Inf. Syst.. IEEE, 2017.
26. Xu, Z. X., Xu, J., Yan, J. J. et al., Analysis of the diagnostic consistency of Chinese medicine specialists in cardiovascular disease cases and syndrome identification based on the relevant feature for each label learning method. Chinese Journal of Integrative Medicine 21(3):217–222, 2015.
27. Shankaracharya, O. D., Mallick, M. et al., Java-based diabetes type 2 prediction tool for better diagnosis. Diabetes Technol. Ther. 14(3): 251–256, 2012.
28. Sun, B., Li, Y., and Zhang, L., The Intelligent System of Cardiovascular Disease Diagnosis Based on Extension Data Mining. Communications in Computer & Information Science 2008:133–140.
29. Lee, H. G., Noh, K., Lee, B. J. et al., Cardiovascular Disease Diagnosis Method by Emerging Patterns. International Conference on Advanced Data Mining & Applications. Berlin: Springer-Verlag, 2006.
30. Peinado, I., Arredondo, M.T., Villalba, E., et al., Patient interaction in homecare systems to treat cardiovascular diseases in the long term. In: International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, pp 308–311, 2009.
31. Khosla A, Cao Y, Lin C C Y, et al. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 183–192.
32. Qian, P., Sun, S., Jiang, Y., Kuan-Hao, S., Ni, T., Wang, S., and Jr, R. F. M., Cross-domain, soft-partition clustering with diversity measure and knowledge reference. Pattern Recogn. 50:155–177, 2016.
33. Qian, P., Zhou, J., Jiang, Y., Liang, F., Zhao, K., Wang, S., Su, K.-H., and Muzic, Jr., R. F., Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes. IEEE Access 6:28594–28610, 2018.
34. Qian, P., Xi, C., Min, X., Jiang, Y., Kuan-Hao, S., Wang, S., and Jr, R. F. M., SSC-EKE: semi-supervised classification with extensive knowledge exploitation. Inf. Sci. 422:51–76, 2018.
35. Du, S. S., Zhao, M. M., Zhang, Y. et al., Screening for differentially expressed proteins relevant to the differential diagnosis of sarcoidosis and tuberculosis. P1oS one 10(9):e0132466, 2015.
36. Miao, F., Cai, Y. P., and Zhang, Y. T., Risk prediction for heart failure incidence within 1-year using clinical and laboratory factors. Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. IEEE, 2014: 1790-1793.