Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.029258

ARTICLE

Machine Learning-Based Decision-Making Mechanism for Risk Assessment of Cardiovascular Disease

Cheng Wang[1], Haoran Zhu[2,*] and Congjun Rao[2,*]

[1]School of Mathematics and Statistics, Hubei University of Education, Wuhan, 430205, China

[2]School of Science, Wuhan University of Technology, Wuhan, 430070, China

*Corresponding Authors: Haoran Zhu. Email: zhuhrwhut@163.com; Congjun Rao. Email: cjrao@foxmail.com

Received: 09 February 2023 Accepted: 04 April 2023 Published: 22 September 2023

ABSTRACT

Cardiovascular disease (CVD) has gradually become one of the main causes of harm to the life and health of residents. Exploring the influencing factors and risk assessment methods of CVD has become a general trend. In this paper, a machine learning-based decision-making mechanism for risk assessment of CVD is designed. In this mechanism, the logistics regression analysis method and factor analysis model are used to select age, obesity degree, blood pressure, blood fat, blood sugar, smoking status, drinking status, and exercise status as the main pathogenic factors of CVD, and an index system of risk assessment for CVD is established. Then, a two-stage model combining K-means cluster analysis and random forest (RF) is proposed to evaluate and predict the risk of CVD, and the predicted results are compared with the methods of Bayesian discrimination, K-means cluster analysis and RF. The results show that the prediction effect of the proposed two-stage model is better than that of the compared methods. Moreover, several suggestions for the government, the medical industry and the public are provided based on the research results.

KEYWORDS

CVD; influencing factors; risk assessment; machine learning; two-stage model

1 Introduction

With the development of the national economy and the improvement of people's living standards, people's diet and lifestyle have changed greatly, which also brings many health problems, including CVD [1]. CVD is a kind of heart or blood vessel disease, also known as circulatory system disease, which is a series of diseases involving the circulatory system such as ischemic heart disease, endocarditis, cardiomyopathy, peripheral artery disease, and so on. CVD is seriously harmful to human health, with high prevalence, high disability rate and high mortality. It has been reported that CVD has become the first cause of death of human beings, with more than 15 million people dying from CVD every year in the world. The number of deaths from CVD will increase by about 6 million in the next 20 years [2,3].

In recent years, although some achievements have been made in the prevention and treatment of CVD, the mortality rate of CVD is still on the rise. Data shows that there are about 290 million

people suffering from CVD, and more than 3 million people die from CVD every year in China [4,5]. According to research analysis, there are more than 300 factors that may lead to CVD, where the more important factors include age, obesity, high blood pressure, high blood sugar, high blood fat, smoking, drinking, lack of physical exercise, and so on. In addition to the risk of death, the high incidence and mortality of CVD also bring great pressure to patients' individuals, families and society [6,7].

Based on the above background, it has become a general trend to explore the main influencing factors of CVD and the methods to evaluate disease risk. This paper collects risk assessment indicators and actual patient data of CVD, combines statistical models with machine learning methods to quantitatively analyze the main influencing factors of CVD, and establishes a risk assessment system to evaluate the disease risk of CVD. Based on the study results, this paper puts forward some suggestions to the government, the medical industry and the public, respectively, so as to arouse people's attention to the prevention of CVD, and then take effective measures to jointly improve the prevention and control of CVD.

The contributions of this paper are as follows:

(1) This paper combines the methods of univariate variance analysis and logistics regression analysis to study whether each index has a significant impact on CVD. The obtained results are consistent, which makes the investigation conclusions more real and reliable.

(2) The factor analysis model is adopted to classify the main influencing factors of CVD, and the main influencing factors are divided into basic physical conditions, "three high" diseases and living habits, so as to establish a risk assessment system of CVD.

(3) A new two-stage model combining K-means cluster analysis with RF is proposed to evaluate and predict patients' CVD risk. The empirical analysis results show that the proposed method in this paper is better than that of Bayesian discrimination, traditional K-means cluster analysis and RF, which also provides a new method for risk assessment of CVD.

The rest of this paper is organized as follows. Section 2 provides the literature review. Section 3 identifies the main pathogenic factors of CVD by using the methods of logistics regression analysis and factor analysis, and then establishes a risk assessment system of CVD. In Section 4, a two-stage model based on K-means cluster analysis and RF is proposed to evaluate the risk of CVD, and then the proposed model is compared with Bayesian discriminant analysis, K-means cluster analysis and RF to test the reliability and rationality of the proposed two-stage model. Section 5 provides some suggestions for improving the current situation of prevention and control for CVD according to the results of empirical analysis. Section 6 concludes the paper.

2 Literature Review

In terms of the influencing factors for CVD, Steinberger et al. [8] collected the sample data set of seven cities in North Carolina from 1999 to 2001, fitted the three-stage Bayesian hierarchical model, and concluded that the higher the concentration of PM2.5, the greater the probability of residents suffering from CVD. Ferrandiz et al. [9] used the Spanish Rapid Inquiry System (RIF) to conduct more precise and systematic exposure measurement and disease data collection. Through correlation analysis, they found that magnesium in drinking water had a stronger protective effect on mortality of CVD than calcium. Through statistical analysis of big data, Kannel et al. [10] found that the relative risk of CVD was significantly increased with the increase of blood glucose and lipid concentrations. Rosenlund et al. [11] systematically analyzed the patient population of a hospital through case-control method, and concluded that gender, drinking habits, family genetic history, residential area, air quality

and other 10 indicators were the most important influencing factors of CVD. Odden et al. [12] concluded that mental state, environmental factors and physical conditions are the most important risk factors for CVD by using the method of factor analysis. Hunt et al. [13] applied Cox proportional risk regression model to explore the relationship between body mass index (BMI) and the risk of CVD in a rural male population, and concluded that the incidence rate of CVD in the obese population was much higher than that in the normal population.

Similarly, Wu et al. [14] adopted one-way variance analysis and multi-factor logistics regression methods to discuss the current situation and influencing factors of CVD in the elderly population, and concluded that age, smoking history, drinking history and other 7 indicators are the important risk factors for CVD in the elderly. He et al. [15] investigated and analyzed the risk factors related to CVD among citizens in Bao'an District of Shenzhen City through Chi-square test, and established a risk assessment system of CVD. Hou [16] used descriptive statistical analysis, Fisher's exact probability method and logistics regression analysis to study the status quo and the main risk factors of CVD among bank employees in Changchun City, and sorted the influence degree of each factor. He et al. [17] analyzed the influencing factors of CVD in elderly maintenance hemodialysis patients in a hospital in Chongqing through single-factor and multi-factor logistics regression combining with controlled experiments, and concluded that obesity and hypertension are the main causes of CVD. Jiang et al. [18] extracted 256 patients with diabetes and kidney disease from a hospital in Shanghai for difference analysis, and concluded that both diabetes and kidney disease would increase the probability of CVD. Wang et al. [19] analyzed the risk factors of atherosclerotic CVD in the elderly population by stepwise regression method, and the results showed that age, sleep quality, genetic history, obesity and other 9 indicators may lead to the occurrence of CVD. In recent years, population-based cohort studies that investigate epidemiological factors such as genetic, clinical, environmental, lifestyle, and socioeconomic have been conducted for risk factor identification of CVD [20].

In terms of risk assessment of CVD, many scholars have proposed some quantitative methods based on statistics and machine learning [6,21–23]. For example, Kannel et al. [24] created the first risk assessment model of CVD based on Framingham Heart Study, and concluded that nine indicators such as hyperglycemia, hyperlipids and alcohol consumption are the important causes of CVD through case analysis. By collecting research data from 12 countries, European scholars created a risk assessment model for CVD applicable to the European population, which predicted the risk of CVD in the next 10 years through four indicators including gender, age, smoking and drinking [25]. In 2008, the World Health Organization (WHO) published the risk prediction chart of CVD in the Prevention and Treatment of CVD, which introduced a new risk assessment model for countries around the world [26]. The American Heart Association and the Society of Cardiovascular and Cerebrovascular Diseases put forward the summary cohort formula based on the large community cohort of multi-region and multi-ethnicity, including the indicators of age, gender and diastolic blood pressure, and achieved good prediction effect [27]. In 2016, through the study of the latest large cohort data, Professor Gu and his team created the Prediction for ASCVD Risk in China (China-PAR) model to evaluate the 10-year and lifetime risk of CVD. The model takes into account important factors such as age, gender, smoking, alcohol consumption, and congenital genetic disease. It provides an effective tool for improving the protection and management of CVD [28].

In terms of machine learning methods, Polaka et al. [29] used decision tree and RF to classify CVD. Dimopoulos et al. [30] compared K Nearest Neighbors (KNN), decision tree and RF algorithm with traditional scoring models of CVD, and concluded that machine learning methods have better evaluation effect on CVD risk. Chen et al. [31] established a CVD prediction model based on eXtreme Gradient Boosting (XGBoost), and summarized the influence rule of each index according to the

research conclusions. Zheng [32] modeled and predicted CVD risk and found that RF, decision tree and logistics regression all have good predictive value, where RF has the best predictive effect. Wang [33] conducted risk assessment on nearly 10 million patients with CVD based on RF and support vector machine, and achieved good evaluation results. Zhang et al. [34] proposed a prediction model of CVD based on feature selection and probabilistic neural network. Compared with other models, this model used fewer features to achieve better prediction results. Liu et al. [35] used the stacking method to combine classification models such as naive Bayes and logistic regression to predict CVD risk, greatly improving the prediction accuracy. Johri et al. [36] proposed a deep learning artificial intelligence framework for multiclass coronary artery disease prediction by using combination of conventional risk factors, carotid ultrasound, and intraplaque neovascularization.

Through the analysis of the above studies, it can be concluded that scholars have achieved a certain harvest in the fields of influencing factors analysis and risk assessment of CVD, but there are still some room for improvement. First of all, the risk assessment system of CVD established in the existing literature is not perfect, and the selected indicators are not systematic. Secondly, although the machine learning methods have been applied in risk prediction of CVD, the selected models and methods are relatively simple and lack of innovation. Therefore, this paper collects sample data through Kaggle website (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset), explores the main influencing factors of CVD through variance analysis, logistics regression analysis and factor analysis, and establishes a risk assessment system of CVD. The effectiveness of machine learning methods in risk assessment of CVD is tested by Bayesian discrimination, and then a new two-stage model combining cluster analysis with RF is established for risk assessment of CVD. Finally, several suggestions are provided based on the research results in order to improve the prevention and control of CVD.

3 Design of Risk Assessment Index System for CVD

In this section, the selected indicators are introduced in detail, and the risk assessment system of CVD is initially established. Then, the selected data are preprocessed, and the single factor analysis of variance is used for preliminary selection of these indicators.

3.1 Preliminary Selection of Indicators for Risk Assessment of CVD

Based on some existing relative work [1–5,37–42], the risk assessment indicators of CVD are selected initially as follows:

(1) Age

The older you get, the more likely you are to develop CVD. Heart muscle cells increase in size with age. This can cause the heart chamber to dilate slightly and the left ventricle to become unfilled, which can lead to heart failure. In addition, with the passage of time, the aging heart and blood vessels will appear subcellular defects. Even if the body maintains a relatively healthy state, it is prone to a decline in cardiovascular function, leading to common problems such as diastolic insufficiency, which increases the risk of CVD [37].

(2) Gender

Relevant studies [3,5] have shown that physiological differences between the sexes may affect the incidence of CVD. The incidence of CVD in females is generally lower than that in males, because female hormones have protective effects on cardiovascular and cerebrovascular diseases. However,

other data show that women have slightly lower rates of CVD before menopause, but no significant difference after menopause.

(3) Obesity degree

Obesity is harmful to health in many aspects, especially cardiovascular system damage is the most serious threat to life. For people with long-term obesity, the pressure on the heart is heavier, and the contraction needs excessive force, which leads to the damage of the intima of the blood vessels, the appearance of arteriosclerosis, and the formation of fatty cardiomyopathy. In addition, the massive accumulation of fat in the abdomen of obese people will also cause the disorder of lipid metabolism and glucose metabolism, aggravate oxidative stress and inflammatory response, and the lipotoxic factors in the body will promote the apoptosis of cardiomyocytes, leading to reduced cardiac function, and then lead to heart disease and abnormal heart rate [38].

(4) Blood pressure

High blood pressure is one of the most concerned health problems in this century, and it is also an important cause of CVD. Because of long-term elevated blood pressure, the burden of the heart will be aggravated by pumping enough blood volume, which is prone to myocardial hypertrophy and atrial enlargement, and can also damage vascular endothelial function, thus leading to coronary atherosclerosis and causing coronary heart disease. In addition, blood pressure disease will lead to peripheral vascular resistance, increase the elasticity and brittleness of arterial wall, and secondary increase of blood viscosity, thus increasing the probability of heart disease [39].

(5) Blood fat

Hyperlipidemia is easy to block blood vessels and stimulate blood vessels. Since the function of human myocardium is orderly, when hyperlipidemia occurs, it may lead to myocardial dysfunction, which may lead to hemorrhagic and apoplexy cerebral thrombosis, thus threatening the safety of patients. In addition, studies have also shown that hyperlipidemia may lead to atherosclerosis, vascular embolism and other diseases [40].

(6) Blood sugar

Excessive blood sugar may lead to vascular wall damage, affect blood flow, and increase blood viscosity. In addition, a study by a foreign collaborative group of emerging risk factors found that diabetes has an extremely serious impact on the risk of death from CVD, which can lead to about twice the probability of death from CVD [41].

(7) Smoking status

Smoking is the leading cause of human death worldwide, with data showing that CVD accounts for 40 per cent of all smoking-related deaths. As an important cause of CVD, smoking increases heart rate, reduces cardiac output and coronary blood flow, leads to short- and long-term increases in blood pressure, and significantly increases the risk of ischemic stroke, which in turn has a significant impact on cardiovascular function.

(8) Drinking status

Drinking alcohol will speed up the heartbeat and blood flow of the human body. Excessive alcohol consumption can cause reduced systolic function of the heart muscle, leading to high blood pressure, stroke and secondary heart disease. In addition, long-term alcohol consumption will affect the metabolism of sugars and fats in the body, resulting in elevated triglycerides, obesity and

hypertriglyceridemia in patients, thus increasing the risk of CVD. Other studies have shown that there is a causal relationship between alcohol consumption and CVD risk, and the trend is basically linear [42].

(9) Exercise status

Regular physical exercise plays a certain role in preventing CVD. It can improve the function of the cardiovascular system, slow down the heart rate and enhance the contractility of the myocardium, thereby increasing the output of the heart pulse and improving the microcirculation. In addition, physical exercise can also promote the dilation of coronary arteries, increase the number of myocardial capillaries, and improve myocardial ischemia.

In summary, 9 factors, i.e., age, gender, obesity degree, blood pressure, blood fat, blood sugar, smoking status, drinking status and exercise status, are preliminarily selected as evaluation indicators for risk assessment of CVD in this paper.

3.2 Preprocessing for Index Values

3.2.1 Data Sources and Standardization

In this paper, the data are selected from Kaggle site survey report (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). After eliminating some abnormal data, we get 68241 valid samples from 70000 patients. According to the evaluation indicators obtained in Section 3.1, the original data of each index are standardized and listed as shown in Table 1.

Table 1: The standardized data

Serial number	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	3	2	1	1	1	1	0	0	1	0
2	3	1	3	1	3	1	0	0	1	1
3	3	1	1	1	3	1	0	0	0	1
4	2	2	2	2	1	1	0	0	1	1
5	2	1	1	1	1	1	0	0	0	0
6	4	1	2	1	2	2	0	0	0	0
7	4	1	3	1	3	1	0	0	1	0
8	4	2	2	1	3	3	0	0	1	1
9	2	1	2	1	1	1	0	0	1	0
10	3	1	2	1	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
68237	3	1	2	2	1	1	0	0	1	1
68238	3	2	2	1	1	1	1	0	1	0
68239	3	2	3	3	3	1	0	1	0	1
68240	4	1	2	1	1	2	0	0	0	1
68241	3	1	1	1	2	1	0	0	1	0

The corresponding indicators of the 9 variables in Table 1 and their meanings are shown in Table 2.

Table 2: Evaluation indicators and their meanings

Indicators	Meanings	Standardized data meaning
X_1	Age	1: 40 years old and younger, 2: 40 to 50 years old, 3: 50 to 60 years old, 4: 60 years old and older
X_2	Gender	1: Female, 2: Male
X_3	Obesity degree	1: Normal, 2: Overweight, 3: Obese
X_4	Blood pressure	1: Normal, 2: Above normal, 3: Much above normal
X_5	Blood fat	1: Normal, 2: Above normal, 3: Much above normal
X_6	Blood sugar	1: Normal, 2: Above normal, 3: Much above normal
X_7	Smoking status	1: Have smoking habit, 0: Have no smoking habit
X_8	Drinking status	1: Have a drinking habit, 0: Have no drinking habit
X_9	Exercise status	1: Have exercise habits, 0: Have no exercise habits
Y	Have CVD or not	1: Sick, 0: Not sick

Note: Obesity degree is measured by body mass index (BMI), BMI= weight (kg)/height^2 (m^2).

3.2.2 Preliminary Screening of Risk Assessment Indicators via Univariate Analysis of Variance

In this section, the univariate analysis of variance [43–45] is used to analyze the independent test results of the 9 indicators, so as to test whether the indicators have a significant impact on CVD, and eliminate the indicators that have a small impact on CVD. SPSS software is used to conduct univariate analysis of variance on standardized data, and the results are shown in Table 3.

Table 3: The results of univariate analysis of variance

Indicators	Classification	Sample number n	Number of diseases	χ^2 value	P value
Age	40 years old and younger	1753	420	259.960	0.000
	40 to 50 years old	19165	7185		
	50 to 60 years old	34639	17707		
	60 years old and older	12684	8472		
Gender	Male	44397	21858	1.203	0.051
	Female	23844	11926		
Obesity degree	Normal	25339	9992	280.239	0.000
	Overweight	25168	12704		
	Obese	17734	11088		
Blood pressure	Normal	57875	24985	765.309	0.000
	Above normal	8384	7088		
	Much above normal	1982	1711		

(Continued)

Table 3 (continued)

Indicators	Classification	Sample number n	Number of diseases	χ^2 value	P value
Blood fat	Normal	51201	22315	417.794	0.000
	Above normal	9221	5507		
	Much above normal	7819	5962		
Blood sugar	Normal	58047	27637	71.072	0.000
	Above normal	5015	2952		
	Much above normal	5179	3195		
Smoking status	Have smoking habit	6007	2821	4.516	0.000
	Have no smoking habit	62234	30963		
Drinking status	Have a drinking habit	3655	1754	1.140	0.000
	Have no drinking habit	64586	32030		
Exercise status	Have exercise habits	54832	26631	24.830	0.000
	Have no exercise habits	13409	7153		

From Table 3, the P value of the significance test of gender is greater than 0.05, which can be considered that gender has no significant effect on CVD. While the P value of the significance test of age, obesity degree, blood pressure, blood fat, blood sugar, smoking status, drinking status and exercise status are all less than 0.05, which means they all pass the significance test. Therefore, 8 factors, i.e., age, obesity degree, blood pressure, blood fat, blood sugar, smoking status, drinking status and exercise status are preliminarily determined as the influencing factors of CVD.

3.3 Identification of Main Pathogenic Factors Using Logistic Regression and Factor Analysis

Based on the 8 influencing factors of CVD and the standardized data obtained in Section 3.2, the method of logistics regression analysis is firstly used to further verify the results of univariate analysis of variance. Secondly, a factor analysis method is applied to classify the evaluation indicators obtained, so as to identify the main pathogenic factors of CVD and establish the final risk assessment index system of CVD.

3.3.1 Test of Influencing Factors Using Logistics Regression Analysis Model

In order to further verify the reasonability of the influencing factors obtained from univariate analysis of variance, a logistics regression analysis model is established, and 8 indicators that have a significant impact are taken as independent variables for testing. Different from the traditional linear regression model, the dependent variables of logistics regression model are discrete, so the independent variable cannot be used directly for the regression of dependent variable, but can be used to regress the probability value of the dependent variable.

The value of the dependent variable y_i (whether to suffer from CVD) in the logistics regression model is only 0 or 1. P is set to represent the probability of $y = 1$ (getting sick), and Q is the probability of $y = 0$ (not sick), which satisfies $Q = 1 - P$. The m influencing factors of CVD are denoted as the vector $X = (x_1, x_2, \ldots, x_m)'$, and the specific relationship between y and X is expressed by

$$P = P(y = 1|X) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_m)} = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \quad (1)$$

By logit transformation for Eq. (1), the final logistics regression model can be expressed by

$$Logit(y) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m = X\beta \quad (2)$$

where $m = 8$, and $\beta_0, \beta_1, \ldots, \beta_m$ are the estimated parameters of the influencing factors of CVD.

The standardized data is substituted into the logistics regression analysis model (2). The software of SPSS is used to solve the model (2), and the results are shown in Table 4.

Table 4: Results of logistics regression analysis

Influencing factors	Coefficient	Standard error	*Wald* χ^2	*P* values	*Exp*(*B*)
Age	0.503	0.012	1845.343	0.000	1.653
Obesity degree	0.295	0.011	737.382	0.000	1.343
Blood pressure	1.489	0.026	3234.701	0.000	4.431
Blood fat	0.589	0.015	1579.315	0.000	1.803
Blood sugar	0.106	0.017	39.040	0.000	0.900
Smoking status	0.607	0.032	4.477	0.034	0.935
Drinking status	0.163	0.040	16.262	0.000	0.850
Exercise status	−0.221	0.021	111.264	0.000	0.802

As can be seen from Table 4, the 8 indicators all have passed the significance test of logistics regression analysis model, indicating the reliability of the results obtained by univariate analysis of variance. Moreover, according to the results of regression coefficients, we can conclude that people with older age, obesity, hypertension, high blood fat, high blood sugar, smoking and drinking habits, and no exercise habits are more likely to suffer from CVD.

3.3.2 Identification of Major Pathogenic Factors of CVD Using Factor Analysis Model

In order to find out the root cause of CVD and formulate reasonable preventive measures, this paper establishes a factor analysis model [46–48]. Based on the mutual relationship of the influencing factors of CVD, the influential factors with high correlation are grouped into one class to obtain a common factor, and finally achieve the effect of dimensionality reduction. According to the above basic ideas, the eight influencing factors are systematically classified through data preprocessing, Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity [49,50], factor extraction and factor rotation, and factor score calculation, so as to form an evaluation system of CVD. The steps of factor analysis model are given as follows:

Step 1: Data preprocessing

According to the results of logistics regression analysis, it can be concluded that the sample data corresponding to the exercise status factor in the data set is negatively correlated with the sample data corresponding to the CVD, while the sample data corresponding to the other factors is positively correlated with the sample data corresponding to the CVD. In order to facilitate the study

of cardiovascular evaluation system, the data set corresponding to exercise status is taken as a negative sign, so that all factors have a positive effect on CVD risk.

Step 2: KMO test and Bartlett's test of sphericity

In order to test whether there is correlation between various influencing factors, and further explain the feasibility of factor analysis, the KMO test and Bartlett's test of sphericity are used on sample data.

The ideas and steps of KMO test [49] are as follows. KMO test statistic is used to compare simple correlation coefficient and partial correlation coefficient between variables. It is mainly used in factor analysis of multivariate statistics. The KMO statistic is between 0 and 1. When the sum of the squares of the simple correlation coefficients among all variables is much larger than the sum of the squares of the partial correlation coefficients, the closer the KMO value is to 1, which means that the correlation between variables is stronger, and the original variables are more suitable for factor analysis. When the sum of squares of the simple correlation coefficients among all variables is close to 0, the closer the KMO value is to 0, which means that the correlation between variables is weaker and the original variables are less suitable for factor analysis.

The ideas and steps of Bartlett's test of sphericity [49,51] are as follows. Bartlett's spherical test is a test method to test the degree of correlation between various variables, which is generally carried out before factor analysis to determine whether variables are suitable for factor analysis. Bartlett's sphericity test is based on the correlation coefficient matrix of variables. Its correlation coefficient matrix of null hypothesis is an identity matrix, that is, all elements on the diagonal of the correlation coefficient matrix are 1 and all non-diagonal elements are zero. The statistics of Bartlett's sphericity test are obtained from the determinant of the matrix of correlation coefficients. If the value is large, and its corresponding associated probability value is less than the significance level in the user's mind, then the null hypothesis should be rejected and the correlation coefficient should be considered as an identity matrix. That is, there is correlation between the original variables, which is suitable for factor analysis. On the contrary, it is not suitable for factor analysis.

By implementing the KMO test and Bartlett's test of sphericity, the test results are shown in Table 5.

Table 5: Results of KMO test and Bartlett's test of sphericity

KMO values		0.553
Bartlett's test of sphericity	Approximate chi-square	31115.613
	Degree of freedom	28
	Significance (*P* value)	0.000

It can be seen from Table 5 that the *P* value of the significance test is less than 0.05, so the null hypothesis is rejected, indicating that there is a correlation between different influencing factors of CVD, so the factor analysis can be conducted.

Step 3: Factor extraction

The 8 impact factors are recombined, the same type of impact factors is inductively combined, and finally several common factors with the greatest explanatory power are selected. The results are shown in Table 6.

Table 6: Feature extraction

Component	Total	Percentage of variance (%)	Cumulative contribution rate (%)
1	2.648	33.1%	33.1%
2	2.354	29.4%	62.5%
3	1.565	19.6%	82.1%
4	0.497	6.2%	88.3%
5	0.421	5.3%	93.6%
6	0.318	4.0%	97.6%
7	0.157	1.9%	99.5%
8	0.041	0.5%	100%

From Table 6, the eigenvalue of the first three factors exceeds 1 and the cumulative contribution rate reaches 82.1%, indicating that these three common factors have been able to cover most of the information of the eight influencing factors of CVD. Therefore, the number of common factors selected by factor analysis in this paper is 3.

Step 4: Factor rotation

In order to make the results easier to understand, this paper uses the maximum variance method to rotate the load matrix and increase the difference between the loads of various factors, which makes the specific meaning of the load matrix clearer, so as to identify the main pathogenic factors of CVD.

SPSS software is used to calculate, and the final factor load matrix is shown in Table 7.

Table 7: Factor load

Factors	Common factor F_1	Common factor F_2	Common factor F_3
X_1	−0.164	−0.171	0.538
X_2	−0.112	−0.012	0.669
X_3	0.761	0.053	−0.052
X_4	0.818	0.051	0.190
X_5	0.859	0.019	0.016
X_6	−0.027	0.809	0.001
X_7	0.001	0.806	0.073
X_8	0.021	−0.623	−0.050

As can be seen from Table 7, common factor F_1 has large loading values on X_3 (blood pressure), X_4 (blood fat) and X_5 (blood sugar), which can be regarded as "three high" disease factors. The loading values of common factor F_2 on X_6 (smoking status), X_7 (drinking status) and X_8 (exercise status) are large, which can be regarded as lifestyle factors. Common factor F_3 has large loading values on X_1 (age) and X_2 (obesity degree), which can be regarded as basic body condition factors.

Step 5: Calculate the factor score

The factor score is estimated by the regression method, and the proportion of variance contribution rate of each common factor in the total variance contribution rate of the three common factors is taken as the weight for weighted summary, i.e.,

$$F = (2.648F_1 + 2.354F_2 + 1.565F_3)/6.567 \tag{3}$$

Therefore, the comprehensive scores of 68241 patients in CVD risk are obtained, and the results are shown in Table 8 (SF1, SF2 and SF3 represent the scores of each sample on the three common factors, respectively).

Table 8: The comprehensive score of each patient

Serial number	SF_1	SF_2	SF_3	Comprehensive score
1	−0.4220	−0.3236	−0.7920	−0.5041
2	1.1612	−0.2665	0.6524	0.5279
3	1.1664	−0.4544	−0.5937	0.0802
4	−0.8651	−0.1014	0.8323	−0.0840
5	−0.5462	−0.3768	−1.1343	−0.6723
6	1.4568	−0.6823	0.2904	0.3822
7	1.2285	−0.4441	1.1099	0.6345
8	3.4132	−0.3802	0.0039	1.0920
9	−0.5198	−0.1675	−0.5689	−0.4177
10	−0.5108	−0.5758	0.0039	−0.3727
...	...	...	...	...
68237	−0.6085	−0.4198	0.2270	−0.2863
68238	0.4007	−0.2736	−0.7505	−0.1811
68239	1.4162	−0.2953	0.3982	0.5301
68240	0.8471	−0.1789	1.3729	0.6686
68241	1.7375	−0.5017	0.6016	0.6391

From Table 8, the comprehensive score of the samples suffering from CVD is higher, while the comprehensive score of the samples without CVD is lower, indicating that the factor score is scientific to a certain extent.

Based on the above analysis, a risk assessment index system of CVD is established, as shown in Fig. 1.

4 Risk Assessment of Two-Stage Model via K-Means Clustering and RF

Based on the standardized data and the final influencing factors of CVD obtained in Section 3, a two-stage model integrating K-means clustering analysis and RF is proposed to evaluate the risk of CVD, and the empirical analysis based on the sample data of 68241 patients in a Kaggle dataset is conducted. Moreover, the evaluation effect obtained by the two-stage model is compared with the traditional methods including Bayesian discrimination, K-means cluster analysis and RF, so as to

verify the reliability of the two-stage model in improving the effect of risk assessment of CVD, and the feasibility of machine learning method in risk assessment of CVD.

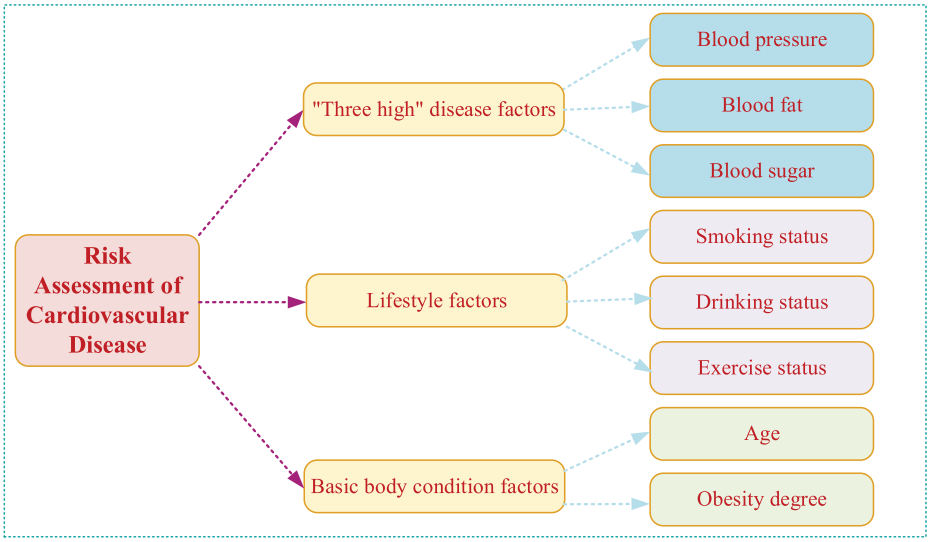

Figure 1: Risk assessment index system of CVD

4.1 K-Means Cluster Analysis

To evaluate the CVD of each patient, 68241 patients could be classified according to the relevant data of each indicator through K-means cluster analysis [52–54]. The main idea is to select K patients in the data set as the initial center of each class according to certain strategies, and then divide the remaining patients into the class closest to the K patients, so as to complete a division. However, the K classes formed are not necessarily the best division, so it is necessary to recalculate the central point of each new class, and then re-divide until the results of each division are consistent. The specific algorithm flow is as follows:

(1) Any K patients from the n patients are selected as the initial clustering center.

(2) Calculate the distance between the remaining patients and the K patients, and divide them into the closest category.

(3) The center of each class is recalculated to obtain K new clustering centers.

(4) Cycle the above Steps (2) and (3) until the results of each partition no longer change.

The standard of stable location of clustering center is: make the error of clustering center change reach a certain threshold or reach a certain number of iterations, and the error sum of squares is:

$$SSE = \sum_{i,j} (x_{ij} - c_i)^2 \tag{4}$$

where c_i represents the i-th classification, $i = 1, 2, \ldots, k$, and x_{ij} represents the indicator data of the j-th patient in the i-th classification.

4.2 Random Forest

In addition to K-means cluster analysis, the RF model can be used to evaluate the patients' CVD risk. This model obtains n sample sets through n random sampling of patient samples. For each sample set, the decision tree model can be trained independently. For the results of n decision tree models, the method of simple majority voting is used to determine the final prediction results [55–57]. It should be noted that the n decision tree models are independent of each other, but not completely independent. There can be intersection between training sets.

Decision tree is the basis of establishing the RF model, which is a top-down tree classification algorithm for patient sample data. It is composed of nodes and directed edges. Nodes are divided into root nodes, child nodes and leaf nodes. The root node represents the first traversed and most important influencing factors of CVD, each child node represents other influencing factors that may be traversed, and leaf node represents the category of patients eventually assigned. Starting from the top root node, all patients are gathered together. After the division of the root node, patients are assigned to different sub-nodes, and then further divided according to the influencing factors in the sub-nodes until all patients are respectively classified into a certain category (namely leaf node) [58,59].

In order to select appropriate influencing factors of CVD and construct decision tree, feature selection is needed. Its function is to select the factors that have a great influence on the decision tree algorithm from the 8 factors affecting CVD, so that the patient samples contained in the leaf nodes of the decision tree belong to the same category as far as possible, even if the nodes are of higher "purity".

Gini coefficient is usually used in decision tree to measure node impurity, which reflects two patients randomly selected from the data set with different probabilities marked by their categories. The calculation formula is as follows:

$$Gini\,(t) = 1 - \sum_{i=0}^{c-1} [p\,(i|t)]^2 \tag{5}$$

where c is the number of classes, t is the node position, $p(i|t)$ is the frequency of the class i distribution, and the Gini coefficient value is in the range of [0, 1].

The implementation process of RF mainly includes the following steps:

(1) c samples are randomly selected from the original 68241 patient samples as the training set, and the remaining samples as the test set.

(2) A new patient data set is obtained by randomly extracting m samples from patient training set C through sampling with retractions.

(3) A certain number of influencing factors are randomly selected from the original eight influencing factors of CVD through sampling without retracting to form a new feature subspace on which a decision tree was established.

(4) Repeat Steps (1) and (2) n times to generate n sample sets $D_1, D_2, \ldots, D_n$ and n decision tree models $T_1, T_2, \ldots, T_n$.

(5) The n decision trees are used to independently classify the test set of patient samples, and the final prediction results are determined based on the classification results combined with the majority voting mechanism [56]. Among them, the classification prediction result of the sample test set can be expressed by Eq. (6):

$$f\,(x) = \arg\max_{Y} \sum_{i=1}^{n} I\,(h_i\,(X) = Y) \tag{6}$$

where h_i is the basic classification model of a single classification tree, and Y is the output variable, namely whether to suffer from CVD.

4.3 Two-Stage Model Integrating K-Means Clustering Analysis and Random Forest

It can be concluded from the above analysis that the risk of CVD can be evaluated by K-means cluster analysis or RF model. However, the K-means clustering analysis is an iterative method, only the local optimal solution can be obtained. In addition, due to the large number of influencing factors in this paper, the RF may reduce the prediction accuracy of the model due to improper feature selection. In order to improve the effect of risk assessment of CVD, a two-stage model that integrates K-means cluster analysis and RF is proposed in order to provide a new method for machine learning in risk assessment of CVD. The specific process is as follows:

(1) A K-means cluster analysis is performed on 68241 patients, and the samples are divided into K categories.

(2) Comparing factor scores, the M group with the highest score is classified as patients, while the M group with the lowest score is classified as non-patients.

(3) The remaining K-2 M sample set is reorganized to obtain a new patient data set, and then the RF model is used for discrimination and classification.

(4) The final risk assessment results of CVD are obtained by combining the results of K-means clustering analysis and RF.

4.4 Empirical Analysis

4.4.1 Initial Classification Based on K-Means Cluster Analysis

Combining with 8 influencing factors, the samples are divided into 4 categories by K-means clustering analysis method. Through repeated iterations with SPSS software, the final 4 clustering centers are obtained, as shown in Table 9.

Table 9: The clustering centers

Indexes	1	2	3	4
Age	3	3	3	2
Obesity degree	3	2	2	1
Blood pressure	1	1	1	1
Blood fat	2	1	1	1
Blood sugar	3	2	1	1
Smoking status	1	0	0	0
Drinking status	1	0	0	0
Exercise status	0	1	1	1

After the initial cluster analysis, 7682 patients are classified as class 1, 26241 as class 2, 8408 as class 3, and 25910 as class 4. Combined the clustering center with the factor scores obtained in Section 3, it can be seen that patients in the first category are generally older, seriously obese, have high blood sugar content, have smoking and drinking habits and lack of exercise, so patients in the first category are judged as patients with CVD. In addition, it can be found that the fourth type of patients are relatively

young, normal physical indicators, no smoking and drinking habits, and often exercise, so the fourth type of patients are judged as non-CVD patients. Among them, the misjudgment matrix of the first and fourth samples is shown in Table 10.

Table 10: False judgment matrix of cluster analysis

Category	Identify as patient (Class 1)	Identify as non-patient (Category 4)
Actual patient (Category 1)	5872 (*TP*)	5606 (*FN*)
Actual non-patient (Category 4)	1810 (*FP*)	20304 (*TN*)

From Table 10, the initial prediction accuracy of K-means clustering is 75.5%. Moreover, it can be found that the CVD risk of patients in the second and third categories obtained by cluster analysis is roughly between the first and fourth categories, which is difficult to distinguish. To further evaluate the CVD risk of these two groups of patients, a total of 34649 patients from the two groups were combined into a new data set, and RF is used for re-prediction.

4.4.2 Reclassification Based on RF

Step 1: Parameter determination

In order to avoid problems such as excessive error rate and poor model fitting, the two most important parameters *mtry* and *ntree* should be determined before establishing the RF model. *mtry* is the number of characteristic variables of the decision tree, and *ntree* is the number of decision tree trees. By using R software, when the number of selected indicators (*mtry*) ranges from 3 to 8, the correctness rate of RF judgment results changes with the increase of *ntree*, as shown in Fig. 2.

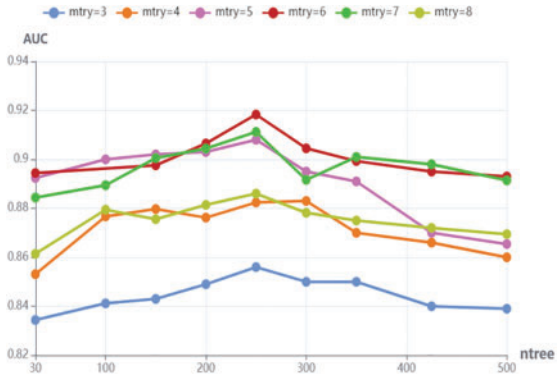

Figure 2: Prediction accuracy with parameter change

According to the results shown in Fig. 2, when the number of randomly selected indicators is 6 and the decision tree is 250 (the maximum value is set), the average decision accuracy rate of the sample reaches the highest, which is 0.916. Thus, we choose $mtry = 6$ and $ntry = 250$.

Step 2: Model implementation

Set $mtry = 6$, $ntry = 250$, integrate the remaining 34649 samples, divide them into test set and training set (sample numbers are 27719 and 6930, respectively) in a ratio of 8:2, and substitute them into RF for judgment. The predicted results are shown in Table 11.

Table 11: Prediction accuracy by RF

Goodness of fit (R^2)	Mean square error ($RMSE$)	Mean absolute error (MAE)	Mean relative error ($MAPE$)
0.8968	1625.2632	623.7428	0.5344

After the prediction, the prediction results of 31073 patients are consistent with the actual situation, and the classification results are obtained by combining K-means clustering analysis. Among 68241 patients, 57249 patients are correctly predicted, and the prediction accuracy of the two-stage model is 83.9%, achieving a good prediction effect. The effectiveness of machine learning method in risk assessment of CVD is further verified.

Step 3: Ranking the importance of indicators

In the stochastic forest model, *IncNodePurity* is an important reference index for ranking the importance of input variables, which can be used as a standard to measure the interpretation degree of characteristic variables to dependent variables. The greater the value, the greater the influence of input variables on the model. By using the *R* software, the importance ranking of the 8 indicators is shown in Table 12.

Table 12: Ranking the importance of indicators

Ranking	Indexes	*IncNodePurity*
1	Blood pressure	1374.16
2	Blood fat	673.24
3	Age	585.24
4	Obesity degree	288.91
5	Blood sugar	93.19
6	Exercise status	50.75
7	Smoking status	33.07
8	Drinking status	30.50

From Table 12, the *IncNodePurity* values of blood pressure, blood fat, age and obesity degree are higher, indicating that these indicators are the most important factors affecting CVD. Among them, high blood pressure, high blood fat, old age, body fat population has a higher probability of CVD. Meanwhile, combined with the results of logistics regression analysis, it is found that the order of regression coefficients is basically consistent with the results of RF, which confirms the reliability of the results of logistics regression model.

4.5 Comparison of Models

In order to further verify the efficiency of the prediction results of the two-stage model combining K-means cluster analysis and RF, this paper compares the two-stage model with the methods of Bayesian discriminant analysis, K-means cluster analysis and RF.

4.5.1 Bayesian Discriminant Analysis

Bayes discriminant analysis [60] is a commonly used discriminant method, which considers prior probability and misjudgment loss, respectively, and takes the maximum probability of an individual belonging to a certain class (or the value of a certain class of discriminant function) or the minimum total average loss of misjudgment as the criterion.

Let $G_1, G_2, \ldots, G_k$ be k populations, and their prior probabilities be $q_1, q_2, \ldots, q_k$, respectively. The density function of each population is denoted as $p_1(x), p_2(x), \ldots, p_k(x)$, respectively, x is a patient sample, and the posterior probability of the patient from the k-th population is:

$$p(j/x) = \frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)} j = 1, 2, \ldots, k \tag{7}$$

If $p(j/x) = \max_{1 \le j \le k} p(j/x)$, then x is judged as the j-th population.

In addition, the concept of minimum misjudgment loss can also be used as a discriminant function. The average loss of x wrongly judged as the j-th population is defined as:

$$E(g/x) = \sum_{j \ne i} \frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)} L(g/j) \tag{8}$$

where $L(g/j)$ is the loss function, which represents the loss of misjudging patients from the j-th population to the g population. When $g = j$, we have $L(g/j) = 0$. When $g \ne j$, $L(g/j) > 0$, and the discriminant criteria are established as follows:

When $E(g/x) = \min_{1 \le j \le k} E(j/x)$, x is judged as the g-th population.

The sample data sets corresponding to the 8 indicators are substituted into the software of SPSS for Bayesian discriminant analysis, and the classification function coefficients are obtained as shown in Table 13.

Table 13: Classification function coefficients

Indicators	Sick	Not sick
X_1 Age	4.777	5.290
X_2 Obesity degree	2.131	2.438
X_3 Blood pressure	4.472	5.680
X_4 Blood fat	0.763	1.337
X_5 Blood sugar	2.440	2.342
X_6 Smoking status	1.394	1.333
X_7 Drinking status	0.040	−0.119
X_8 Exercise status	5.127	4.906
Constant	−15.413	−19.358

Combined with Table 13, two types of linear discriminant functions can be obtained:

Normal group:

$$Y_1 = 4.777x_1 + 2.131x_2 + 4.472x_3 + 0.763x_4 + 2.440x_5 + 1.394x_6 + 0.040x_7 + 5.127x_8 - 15.413 \quad (9)$$

Diseased group:

$$Y_2 = 5.290x_1 + 2.438x_2 + 5.680x_3 + 1.337x_4 + 2.342x_5 + 1.333x_6 - 0.119x_7 + 4.906x_8 - 19.358 \quad (10)$$

The patient sample data is substituted into the linear discriminant function for discrimination, and the discriminant results of 68241 patients are obtained, as shown in Table 14.

Table 14: Bayesian discriminant results

Serial number	Discriminant results	Actual result
1	0	0
2	1	1
3	1	1
4	1	1
5	0	0
6	1	0
7	1	0
8	1	1
9	0	0
10	0	0
...	...	...
68237	1	1
68238	0	0
68239	1	1
68240	1	1
68241	0	0

Note: 1 means having CVD, 0 means not having CVD.

After sorting out the discriminant data, the wrong judgment matrix can be obtained, as shown in Table 15.

Table 15: The wrong judgment matrix

Category	Identify as a patient	Identify as a non-patient
Actual patient	21712 (*TP*)	12072 (*FN*)
Actual nonpatient	7127 (*FP*)	27330 (*TN*)

According to the misjudgment matrix in Table 15, among the actual 34457 patients with non-CVD, 27330 are correctly identified and 7127 are wrongly identified as patients with CVD. While among the actual 33784 patients with CVD, 21712 are correctly identified and 12072 are wrongly identified as patients with non-CVD. To sum up, the accuracy of Bayes discrimination is

$$A = \frac{TP + TN}{TP + FP + TN + FN} = \frac{21712 + 27330}{68241} = 71.9\% \tag{11}$$

The calculation results show that the prediction accuracy of Bayes discriminant method is 71.9%, which means the prediction effect is mediocre.

4.5.2 Comparison of Experiment Results

In this paper, accuracy (*acc*), precision (*pre*), recall (*rec*) and *F*1-*score* are adopted to evaluate the accuracy of model classification [56,61–63]. The specific calculation formulas are as follows:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$pre = \frac{TP}{TP + FP} \tag{13}$$

$$rec = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - score = 2 * \frac{pre + rec}{pre + rec} \tag{15}$$

where *TP*, *TN*, *FP* and *FN* represent the number of true patients, true non-patients, false patients and false non-patients, respectively.

The accuracy, precision, recall and *F*1-*score* of the two-stage model, Bayesian discriminant analysis, K-means clustering analysis and RF are calculated, respectively, and the results are shown in Table 16.

Table 16: Comparison results of models

Models	*acc*	*pre*	*rec*	*F*1-*score*
Two-stage model	0.8393	0.8551	0.8011	0.8272
Bayesian discriminant analysis	0.7186	0.7242	0.6958	0.7103
K-means clustering analysis	0.6876	0.7077	0.6464	0.6757
RF	0.7602	0.7861	0.7425	0.7637

From Table 16, the accuracy rate, precision rate, recall rate and *F*1-*score* of the two-stage model are superior to Bayesian discriminant analysis, traditional K-means clustering analysis and RF model. Thus, the two-stage model combining K-means cluster analysis with RF can effectively improve the effect of predicting CVD risk, which also provides a new method for evaluating CVD risk. In addition, Bayesian discriminant analysis, K-means cluster analysis and RF model also achieve good prediction results, which fully demonstrates the scientific nature of machine learning methods in risk assessment of CVD.

4.6 Test and Improvement of Two-Stage Model

In Section 4.4, $K=4$ and $M=1$ are set, respectively, and the samples are divided into four categories by K-means clustering analysis method. The first category with the highest factor score is classified as CVD patients, and the first category with the lowest factor score is classified as non-CVD patients. The remaining two categories are further evaluated by RF model, which achieved good prediction effect. In order to further test the effectiveness of the two-stage model and explore a scheme to make the prediction results more accurate, different parameters are set for K and M. The results of the prediction accuracy changing with the parameters are shown in Table 17.

Table 17: Variation of prediction accuracy with parameters in two-stage model

K	M	Accuracy rate
4	1	83.9%
5	1	79.6%
5	2	82.4%
6	1	86.4%
6	2	81.7%
7	1	81.2%
7	2	84.2%
7	3	82.6%
8	1	85.8%
8	2	84.8%
8	3	80.4%

As can be seen from Table 17, when $K=6$ and $M=1$, the prediction accuracy of the two-stage model reaches the highest 86.4%. Therefore, the operation of the two-stage model is modified as follows: First, the sample set is divided into 6 categories by K-means clustering analysis method, and the category 1 with the highest factor score is classified as patients with CVD, while the category 1 with the lowest factor score is classified as patients with non-CVD. Secondly, the remaining 4 types of samples are integrated and the RF model is used for re-prediction.

By using the softwares of SPSS and R, the prediction accuracy of single K-means clustering analysis and RF is 68.8% and 76.0%, respectively. Compared with the results in above Table 17, it is found that the prediction accuracy of two-stage model is better than that of single K-means clustering analysis, RF and Bayesian discrimination.

5 Result Analysis and Suggestions

In Sections 3 and 4, the main influencing factors and risk assessment of CVD are studied respectively. In terms of influencing factors of CVD, the univariate analysis of variance and logistics regression analysis are made firstly. The results show that the factors of age, obesity degree, blood pressure, blood fat, blood sugar, smoking status, drinking status and exercise status are the influencing factors of CVD, among which older age, obesity, hypertension, hyperlipidemia, high blood glucose, smoking and drinking habits, no exercise habits of people are more prone to CVD. Secondly, combined with the regression coefficient and RF results, it was concluded that the degree of influence on the risk of CVD from large to small is blood pressure, blood fat, age, obesity degree, blood sugar,

exercise status, smoking status and drinking status. Finally, through the factor analysis model, age and obesity degree are classified as basic physical condition factors, blood pressure, blood sugar and blood fat as "three high" disease factors, smoking status, drinking status and exercise status as lifestyle factors, and the final risk assessment system of CVD is obtained. These conclusions on influencing factors of CVD we obtained by quantitative method is consistent with the cognition of the causative factors of CVD in our daily life. In terms of risk assessment for CVD, combined with 8 influencing factors obtained by Section 3, the risk of CVD of samples is evaluated and predicted by a new two-stage model, and it is concluded that the proposed two-stage model achieved better prediction effect then that of Bayesian discrimination, K-means clustering analysis, and RF, which also provides a new method for risk assessment of CVD.

Based on the above research results, the countermeasures and suggestions on how to scientifically prevent CAD are put forward from the three levels of the government, the medical industry and the public.

(1) From the perspective of government

According to the study results in this paper, smoking, drinking and lack of physical exercise are the main causes of CVD. In view of the above results, the government needs to take targeted measures to reduce the impact of these factors on CVD.

First, make great efforts to popularize health education and improve the health quality of the whole people. It can be seen from the results of logistics regression analysis in Table 4 that the multiple factors such as physical conditions and living habits can lead to the occurrence of CVD. Therefore, the government needs to establish a sound health education system, carry out different forms of CVD publicity and education through media channels, vigorously popularize the harm and prevention knowledge of CVD, enhance the public's awareness and understanding of CVD, and then prevent diseases by improving their lifestyle.

Second, improve the policy environment and people's living habits. The results of logistics regression analysis in Table 4 show that smoking and drinking will cause the occurrence of CVD. Therefore, the government should introduce policies to strengthen the control of smoking in public places and increase the punishment for smoking in public places. At the same time, tax policies on tobacco and alcohol should be improved, prices of tobacco and alcohol should be appropriately raised, and harmful smoking and alcohol consumption should be reduced.

Third, build a healthy living environment and promote the development of national sports. As can be seen from Table 12, lack of physical exercise plays an important role in the causes of CVD. In order to promote the public to carry out sports, the government can build fitness trails, health theme parks and other sports facilities, use the existing resources to create a good environment for sports, so that more people have the opportunity to participate in sports, so as to better implement the policy of national sports.

Fourth, use science and technology to accelerate the integration of emerging medical and Internet achievements. According to the empirical analysis results in Section 4.4, the two-stage model based on machine learning proposed in this paper has good decision support for improving the evaluation effect of CVD. Therefore, it is necessary for the government to make full use of the current advanced science and technology, promote the deep integration of mobile Internet, cloud computing, big data and medical industry, and make full use of emerging technologies to prevent CVD.

(2) From the perspective of the medical industry

According to the results of this study, hypertension, hyperlipidemia, hyperglycemia, obesity and other problems will increase the probability of CVD in the public. Therefore, the medical industry should be targeted to these groups of key protection.

First, implement community-level prevention and control and intensify monitoring of high-risk groups. According to the empirical analysis results in Section 4.4, the machine learning algorithm is scientific in predicting the risk of CVD. Therefore, the medical industry should reach into urban communities and rural areas and set up pilot public health services. In addition, collect public health data, establish health records, predict high-risk groups prone to CVD based on risk factors that may cause CVD, conduct targeted health publicity and education for high-risk groups, provide specific guidance on preventive measures, and urge them to maintain reasonable living habits, so as to reduce the level of risk factors. To achieve the prevention of CVD.

Second, select professionals and set up prevention and control teams. According to the empirical analysis results in Section 4.4, the two-stage model proposed in this paper can effectively improve the prediction accuracy of CVD. Therefore, the medical industry can select outstanding personnel in the industry, set up a CVD prevention and control team, strengthen professional knowledge training and necessary practical training, and enhance their ability to predict the risk of CVD through the investigation and analysis of high-risk groups of CVD, so as to reserve talents for the prevention and control of CVD.

Third, expand consultation services and carry out personalized health intervention. As can be seen from Fig. 1, living habits, "three high" diseases and basic physical conditions are the main pathogenic factors of CVD. Therefore, it is necessary for medical institutions to open counseling services for smoking cessation and alcohol abstinence to advocate the public to quit smoking and alcohol abstinence and maintain good living habits. Meanwhile, combined with the risk assessment method proposed in this paper, the risk assessment and intervention guidance can be gradually carried out for patients with high-risk groups such as obesity, hypertension, hyperglycemia and hyperlipidemia, and consultation services such as reasonable diet, fitness and disease prevention can be provided.

Fourth, deepen the care institutions and promote the integrated development of medical and nursing care. The results of logistics regression analysis in Table 4 show that the elderly is an important group of people suffering from CVD. Therefore, medical institutions need to actively cooperate with pension institutions to establish and improve telemedicine service stations for all the elderly. In addition, community-level medical institutions are encouraged to establish a relationship of assistance with the families of the elderly, and provide medical services such as regular physical check-ups, management of health records and popularizing methods for keeping fit, so as to provide regular medical assistance to the elderly.

(3) From the perspective of the public

The results in this paper showed that the risk factors of CVD mainly came from basic physical conditions, "three high" diseases and poor living habits. Therefore, the public should improve their living habits and pay attention to physical check-ups.

First, eat a reasonable diet to avoid obesity. The results of logistics regression analysis in Table 4 show that obesity is an important factor causing CVD. Therefore, the staple food of the general public should be cereals, and a certain number of fresh vegetables and fruits should be consumed every day. At the same time, the public should try to avoid or eat less fatty meat, animal oil and animal offal, eat

more legumes and put an end to overeating. In addition, the public can drink a small amount of fresh milk every day, but try to avoid some milk products with high oil content.

Second, take care of health by having regular physical examination. Table 12 shows that hypertension, hyperlipidemia and hyperglycemia are the most important risk factors for CVD. Therefore, the general public, especially the elderly, need to go to the hospital for regular physical examination, especially to pay attention to the condition of blood pressure, blood fat and blood sugar. If a certain indicator is too high, the patients need to pay enough attention to it. It is best to have physical therapy under the guidance of a doctor to reduce the risk of developing CVD.

Third, give up smoking and drinking and keep healthy. As can be seen from Table 4, both smoking and drinking are likely to cause CVD. Therefore, it is necessary for the public to strengthen the study of medical and health, so as to increase the awareness of the harm of tobacco and alcohol to CVD, do not smoke, do not drink, so as to maintain good health.

Fourth, insist on exercise to enhance physical fitness. As can be seen from Table 4, moderate physical exercise can reduce the probability of CVD to some extent. Therefore, the public can use their spare time to go to parks, playgrounds, gyms and other places for physical exercise, which should ensure more than 30 min a day, but should not be too long. Through the accumulation of long-term exercise, you can improve your physical fitness, thereby improving immune function and significantly reducing the risk of CVD.

6 Conclusions

With the prevention and control of CVD as the research background, this paper firstly collected patient data through Kaggle platform, selected evaluation indicators of CVD on the basis of reading a large number of literatures, and initially established a risk assessment system of CVD. Then, the univariate analysis of variance and logistics regression analysis are used to test the impact of each indicator on CVD. Combined with the factor analysis model, the indicators that passed the significance test are classified to obtain the final risk assessment system of CVD. Secondly, a new two-stage model integrating K-means clustering analysis and RF is proposed to evaluate the risk of CVD, and the comparative analysis is made to verify the effectiveness and superiority of the proposed two-stage model. Finally, based on the results of empirical analysis, several suggestions are provided for the government, the medical industry and the public to jointly improve the prevention and control of CVD.

The work in this paper still has several limitations, for example, the final risk assessment system of CVD only includes basic physical conditions, "three high" diseases and living habits, while CVD may also be affected by other factors, such as environment. In addition, this paper evaluated and predicted the risk of CVD in samples through the two-stage model combining K-means clustering analysis and RF, but it only conducted the cases when $K = 4$, 6, 8, and lacked comprehensiveness in parameter selection. From these limitations, in order to make the risk assessment results of CVD more valuable, future work should make in the following aspects: (i) In the future investigation of influencing factors of CVD, a more complete evaluation system should be established and more comprehensive factors should be quantitatively analyzed. (ii) More parameters can be added to the parameters of the two-stage model in future work, so as to find the combination methods with the best prediction effect. (iii) It is also our future research direction to propose some novel integrated models based on the combination of deep learning and statistical methods for CVD risk assessment.

Acknowledgement: The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Nos. 72071150, 71871174).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: C. Wang, H. Zhu, and C. Rao; data collection: H. Zhu; analysis and interpretation of results: C. Wang and C. Rao; draft manuscript preparation: H. Zhu and C. Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and analyzed during the current study are available in the [Kaggle site survey report] repository, [https://www.kaggle.com/sulianova/cardiovascular-disease-dataset].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Gaidai, O., Cao, Y., Stas Loginov, S. (2023). Global CVD death rate prediction. *Current Problems in Cardiology, 48(5),* 101622. https://doi.org/10.1016/j.cpcardiol.2023.101622
2. Tawfiq, E., Selak, V., Elwood, J. M., Pylypchuk, R., Tin, S. T. et al. (2023). Performance of CVD risk prediction equations in more than 14000 survivors of cancer in New Zealand primary care: A validation study. *The Lancet, 401(10374),* 357–365. https://doi.org/10.1016/S0140-6736(22)02405-9
3. Wei, X., Rao, C. J., Xiao, X. P., Chen, L., Goh, M. (2023). Risk assessment of CVD based on SOLSSA-CatBoost model. *Expert Systems with Applications, 219(1),* 119648.
4. The Writing Committee of the Report on Cardiovascular Health and Diseases in China (2022). Report on cardiovascular health and diseases in China 2021: An updated summary. *Biomedical and Environmental Sciences, 35(7),* 573–603.
5. Wang, J., Rao, C. J., Goh, M., Xiao, X. P. (2023). Risk assessment of coronary heart disease based on cloud-random forest. *Artificial Intelligence Review, 56(1),* 203–232. https://doi.org/10.1007/s10462-022-10170-z
6. Huang, Y., Ren, Y. B., Yang, H., Ding, Y. J., Liu, Y. et al. (2022). Using a machine learning-based risk prediction model to analyze the coronary artery calcification score and predict coronary heart disease and risk assessment. *Computers in Biology and Medicine, 151(B),* 106297. https://doi.org/10.1016/j.compbiomed.2022.106297
7. Cao, X., Zhao, Z. P., Kang, Y. T., Tian, Y. X., Song, Y. X. et al. (2022). The burden of CVD attributable to high systolic blood pressure across China, 2005–18: A population-based study. *The Lancet Public Health, 7(12),* e1027–e1040. https://doi.org/10.1016/S2468-2667(22)00232-8
8. Steinberger, J., Jacobs, D. R., Raatz, S., Moran, A., Hong, C. P. et al. (2006). Erratum: Comparison of body fatness measurements by BMI and skinfolds vs dual energy X-ray absorptiometry and their relation to cardiovascular risk factors in adolescents. *International Journal of Obesity, 30(7),* 16–18. https://doi.org/10.1038/sj.ijo.0803362
9. Ferrandiz, J., Lopez-Quilez, A., Gomez-Rubio, V., Sanmartin, P., Martinez-Beneito, M. A. et al. (2003). Statistical relationship between hardness of drinking water and cerebrovascular mortality in Valencia: A comparison of spatiotemporal models. *Environmetrics, 14(5),* 491–510. https://doi.org/10.1002/env.601
10. Kannel, W. B., Schwartz, M. J. (2017). The J-curve relationship of treated diastolic blood pressure to mortality risk: Is it real? Is it clinically meaningful? *Current Cardiovascular Risk Reports, 1(3),* 28–31.
11. Rosenlund, M., Njoroge, P., Nisha, I. (2020). Understanding health disparities in CVD in pregnancy among black women: Prevalence, preventive care, and peripartum support networks. *Current Cardiovascular Risk Reports, 14(5),* 57–58.

12. Odden, M., Shlipak, M., Whitson, H., Katz, R., Kearney, P. et al. (2019). Risk factors for CVD across the spectrum of older age: The cardiovascular health study. *Atherosclerosis, 237(1),* 336–342. https://doi.org/10.1016/j.atherosclerosis.2014.09.012
13. Hunt, K., Resendez, R., Williams, K. (2020). National choleateral education program versus world health organization metabolic syndrome in relation to all-cause and cardiovascular mortality in the san-antonio heart study. *Circulation, 110(10),* 1251–1257. https://doi.org/10.1161/01.CIR.0000140762.04598.F9
14. Wu , J., Wang, L., Xie, H. (2021). Status and influencing factors of CVD in elderly population. *Chinese Journal of Public Health Engineering, 20(4),* 599–600.
15. He, X., Liu, Z., Lv, Y. (2021). CVD risk assessment and related risk factors distribution in the elderly in Bao'an district, Shenzhen. *Practical Preventive Medicine, 28(10),* 1153–1157.
16. Hou, D. D. (2017). *Current status and influencing factors of CVD among bank employees in Changchun City*. Jilin, China: Jilin University.
17. He, H., Yang, Y. L. (2020). Influencing factors of CVD in elderly maintenance hemodialysis patients. *Chinese and Foreign Medical Research, 18(19),* 43–46.
18. Jiang, X., Zhou, R., Liu, Y., Shen, J., Wang, Y. N. (2017). Prevalence of CVD and its influencing factors in diabetic nephropathy patients. *Chinese General Practice, 20(29),* 3590–3595.
19. Wang, L., Wu, X., Gu, Y. H., Hu, X. Y., Sha, J. M. et al. (2022). Correlation analysis of factors influencing atherosclerotic CVD in elderly community population. *Journal of Southeast University (Medical Edition), 41(1),* 88–95.
20. Moon, J., Posada-Quintero, H. F., Chon, K. H. (2023). A literature embedding model for CVD prediction using risk factors, symptoms, and genotype information. *Expert Systems with Applications, 213(A),* 118930.
21. Ma, Y. J., Xiong, J. H., Zhu, Y. D. (2022). Deep learning algorithm using fundus photographs for 10-year risk assessment of ischemic CVD in China. *Science Bulletin, 67(1),* 17–20. https://doi.org/10.1016/j.scib.2021.08.016
22. Suri, J. S., Bhagawati, M., Paul, S., Protogeron, A., Sfikakis, P. P. et al. (2022). Understanding the bias in machine learning systems for CVD risk assessment: The first of its kind review. *Computers in Biology and Medicine, 142(1),* 105204.
23. Roseiro, M., Henriques, J., Paredes, S., Rocha, T., Sousa, J. (2023). An interpretable machine learning approach to estimate the influence of inflammation biomarkers on cardiovascular risk assessment. *Computer Methods and Programs in Biomedicine, 230(1),* 107347. https://doi.org/10.1016/j.cmpb.2023.107347
24. Kannel, W. B., Mcgee, D., Gordon, T. (1976). A general cardiovascular risk profile: The framinghan study. *American Journal of Cardiology, 38(1),* 46–47. https://doi.org/10.1016/0002-9149(76)90061-8
25. Conroy, R. (2003). Estimation of ten-year risk of fatal CVD in Europe: The score project. *European Heart Journal, 24(11),* 987–1003. https://doi.org/10.1016/S0195-668X(03)00114-3
26. Organization, W. H. (2007). *Prevention of CVD: Guidelines for assessment and management of cardiovascular risk*. Geneva: World Health Organization.
27. Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. et al. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation, 129(25),* S49–S73.
28. Liu, X. C., Feng, Y. Q., Chen, J. Y. (2021). Research progress of CVD risk assessment model. *Chinese Journal of Practical Internal Medicine, 41(5),* 428–433.
29. Polaka, I., Tom, I., Borisov, A. (2010). Decision tree classifiers in bioinformatics. *Scientific Journal of Riga Technical University Computer Sciences, 42(1),* 118–123. https://doi.org/10.2478/v10143-010-0052-4
30. Dimopoulos, A. C., Nikolaidou, M., Caballero, F. F., Engchuan, W., Sanchez-Niubo, A. et al. (2018). Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BioMed Central, 18(1),* 262–265. https://doi.org/10.1186/s12874-018-0644-1
31. Chen, D. T., Chan, J., Zhou, W. D. (2021). Prediction model and index analysis of CVD based on XGboost. *Modern Hospitals, 21(6),* 958–961.

32. Zheng, X. Y. (2018). *CVD prediction system based on machine learning*. Beijing, China: Beijing Jiaotong University.
33. Wang, J. X. (2018). *Prediction of CVD based on random forest and support vector machine*. Tianjin, China: Tianjin University.
34. Zhang, Z. H., Hu, W. P. (2022). Prediction of heart disease based on feature selection and probabilistic neural networks. *Modern Electronic Technology, 45(1),* 95–99.
35. Liu, Y., Fu, J. F., Hu, J. K. (2021). Prediction of CVD based on classification model. *Computer Programming Skills & Maintenance, 23(11),* 23–24.
36. Johri, A. M., Singh, K. V., Mantella, L. E., Saba, L., Sharma, A. et al. (2022). Deep learning artificial intelligence framework for multiclass coronary artery disease prediction using combination of conventional risk factors, carotid ultrasound, and intraplaque neovascularization. *Computers in Biology and Medicine, 150(1),* 106018. https://doi.org/10.1016/j.compbiomed.2022.106018
37. Rossouw, J. E., Prentice, R. L., Manson, J. E., Liu, S. W. (2007). Effect of age and menopause years on the association between postmenopausal hormone therapy and CVD risk. *Dkgest of the World Latest Medical Information, 9(1),* 2–4.
38. Bianchettin, R. G., Lavie, C. J., Lopez-Jimenez, F. (2023). Challenges in cardiovascular evaluation and management of obese patients: JACC state-of-the-art review. *Journal of the American College of Cardiology, 81(5),* 490–504. https://doi.org/10.1016/j.jacc.2022.11.031
39. Nazarzadeh, M., Bidel, Z., Canoy, D., Copland, E., Bennett, D. A. et al. (2022). Blood pressure-lowering treatment for prevention of major CVD in people with and without type 2 diabetes: An individual participant-level data meta-analysis. *The Lancet Diabetes & Endocrinology, 10(9),* 645–654. https://doi.org/10.1016/S2213-8587(22)00172-3
40. Yang, X. J., Li, Q. M., Liu, D. C., Han, M. H., Qie, R. R. et al. (2022). Joint effect of physical activity and blood lipid levels on all-cause and CVD mortality: The rural Chinese cohort. *Nutrition Metabolism and CVD, 32(6),* 1445–1453.
41. Schwarz, P., Timpel, P., Harst, L., Greaves, C. J., Ali, M. K. et al. (2018). Blood sugar regulation for cardiovascular health promotion and disease prevention: JACC health promotion series. *Journal of the American College of Cardiology, 72(15),* 1829–1844. https://doi.org/10.1016/j.jacc.2018.07.081
42. Medling, T., Gobeil, K., Sawalha, K., Abozenah, M., Tavares, P. et al. (2022). Relation of patient's opinion of alcohol's health effects and drinking habits among hospitalized patients with CVD. *The American Journal of Cardiology, 179(1),* 31–38.
43. Qu, S. J., Feng, C. (2022). Data-driven robust DEA models for measuring operational efficiency of endowment insurance system of different provinces in China. *Sustainability, 14(16),* 9954. https://doi.org/10.3390/su14169954
44. Rao, C. J., Zhang, Y., Wen, J. H., Xiao, X. P., Goh, M. (2023). Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model. *Energy, 263(C),* 125955. https://doi.org/10.1016/j.energy.2022.125955
45. Ding, Q., Xiao, X. P., Kong, D. K. (2023). Estimating energy-related CO_2 emissions using a novel multivariable fuzzy grey model with time-delay and interaction effect characteristics. *Energy, 263(E),* 126005. https://doi.org/10.1016/j.energy.2022.126005
46. Ahmadian, A., Sedghi, M., Fgaier, H. (2019). PEVs data mining based on factor analysis method for energy storage and DG planning in active distribution network: Introducing S2S effect. *Energy, 175(1),* 265–277. https://doi.org/10.1016/j.energy.2019.03.097
47. Chu, J. J., Xiao, X. P. (2023). Benefits evaluation of the northeast passage based on grey relational degree of discrete Z-numbers. *Information Sciences, 626,* 607–625. https://doi.org/10.1016/j.ins.2023.02.085
48. Gao, M. Y., Yang, H. L., Xiao, Q. Z., Goh, M. (2022). A novel method for carbon emission forecasting based on Gompertz's law and fractional grey model: Evidence from American industrial sector. *Renewable Energy, 181(1),* 803–819. https://doi.org/10.1016/j.renene.2021.09.072

49. Wu, T. (2021). Quantifying coastal flood vulnerability for climate adaptation policy using principal component analysis. *Ecological Indicators, 129(1),* 108006. https://doi.org/10.1016/j.ecolind.2021.108006
50. Chen, L., Dong, T., Nan, G. F., Xiao, Q. Z., Xu, M. et al. (2023). Impact of the introduction of marketplace channel on e-tailer's logistics service strategy. *Managerial and Decision Economics, 44(5),* 2835–2855. https://doi.org/10.1002/mde.3850
51. Yin, C., Mao, S. H. (2023). Fractional multivariate grey Bernoulli model combined with improved grey wolf algorithm: Application in short-term power load forecasting. *Energy, 269(1),* 126844. https://doi.org/10.1016/j.energy.2023.126844
52. Wei, J. P., Qu, S. J. (2022). The novel data-driven robust maximum expert mixed integer consensus models under multi-role's opinions uncertainty by considering non-cooperators. *IEEE Transaction on Computational Social Systems*. https://doi.org/10.1109/TCSS.2022.3192897
53. Li, Y., Chu, X. Q., Tian, D., Feng, J. Y., Mu, W. S. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing, 113(B),* 107924. https://doi.org/10.1016/j.asoc.2021.107924
54. Peng, J. J., Chen, X. G., Tian, C., Zhang, Z. Q., Song, H. Y. et al. (2022). Picture fuzzy large-scale group decision-making in a trust-relationship-based social network environment. *Information Sciences, 608(1),* 1675–1701. https://doi.org/10.1016/j.ins.2022.07.019
55. Qu, S. J., Xu, L., Mangla, S. K., Chan, F., Zhu, J. L. et al. (2022). Matchmaking in reward-based crowdfunding platforms: A hybrid machine learning approach. *International Journal of Production Research, 60(24)*. https://doi.org/10.1080/00207543.2022.2121870
56. Rao, C. J., Liu, M., Goh, M., Wen, J. H. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. *Applied Soft Computing, 95(1),* 106570.
57. Zhang, Y. X., Guo, H., Sun, M., Liu, S. F., Forrest, J. (2023). A novel grey Lotka–Volterra model driven by the mechanism of competition and cooperation for energy consumption forecasting. *Energy, 264(1),* 126154. https://doi.org/10.1016/j.energy.2022.126154
58. Rao, C. J., Liu, Y., Goh, M. (2022). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model. *Complex & Intelligent Systems, 9,* 1391–1414. https://doi.org/10.1007/s40747-022-00854-y
59. Wen, J. H., Hong, L. J., Dai, M., Xiao, X. P., Wu, C. Z. (2023). A stochastic model for stop-and-go phenomenon in traffic oscillation: On the prospective of macro and micro traffic flow. *Applied Mathematics and Computation, 440(1),* 127637. https://doi.org/10.1016/j.amc.2022.127637
60. Wang, J. X., Zhang, X. Q., Chang, X. K., Zhou, Z. H., Wang, C. et al. (2022). Prediction model for blast-induced structural damage based on an optimized Bayes discriminant analysis model. *Applied Mathematical Modelling, 110(1),* 338–366. https://doi.org/10.1016/j.apm.2022.05.011
61. Rao, C. J., Gao, M. Y., Wen, J. H., Goh, M. (2022). Multi-attribute group decision making method with dual comprehensive clouds under information environment of dual uncertain Z-numbers. *Information Sciences, 602(1),* 106–127. https://doi.org/10.1016/j.ins.2022.04.031
62. Qu, S. J., Shu, L. L., Yao, J. Y. (2022). Optimal pricing and service level in supply chain considering misreport behavior and fairness concern. *Computers & Industrial Engineering, 174,* 108759. https://doi.org/10.1016/j.cie.2022.108759
63. Rao, C. J., Wang, C., Hu, Z., Xiao, X. P., Goh, M. (2023). Gray uncertain linguistic multiattribute group decision making method based on GCC-HCD. *IEEE Transactions on Computational Social Systems, 10(2),* 523–537. https://doi.org/10.1109/TCSS.2022.3166526