Asia Pacific Science Press

Asia Pacific Journal of Clinical Medical Research
ISSN(O):3079-8345
ISSN(P):3079-8337
Vol. 1 No.1 (2025)

Stroke Risk Prediction and Assessment Based on Big Data Analysis

Menghan Gao*, Jiayin Chen, Hongyan Gao

School of Management, Tianjin University of Technology,Tianjin 300384,China

Corresponding author: Menghan Gao, gaomenghan123456@gmail.com

Abstract: Stroke is a common cardiovascular and cerebrovascular disease with high morbidity, high mortality and high disability rate. In this paper, a stroke risk prediction and evaluation model based on support vector machine, random forest, BP neural network and genetic algorithm optimization neural network algorithm was established by using a raw dataset including 10 characteristic variables such as gender, age, hypertension, heart disease, and 1 stroke target variable. The experimental results show that the average blood glucose level, body mass index, hypertension and other variables have a great impact on the risk of stroke, and the neural network algorithm optimized by the genetic algorithm performs slightly better than the other three models.

Keywords: Stroke; BP Neural Network; Genetic Algorithm; Support Vector Machine; Random Forest; Risk Prediction Evaluation

Published: Mar 24, 2025

DOI: https://doi.org/10.62177/apjcmr.v1i1.196

1.Introduction

Cerebral apoplexy, commonly known as stroke. The latest Global Burden of Disease study (GBD) shows that the overall lifetime risk of stroke in China is 39.9%, ranking first in the world, almost 2 out of every 5 people suffer from stroke, and about 1.94 million people die from stroke in China every year. According to the data of the Stroke Prevention and Control Engineering Committee of the National Health Commission, in 2020, the prevalence rate of stroke in residents over 40 years old in China was 2.61%, the incidence rate was 505.23/100,000, and the mortality rate was 343.4/100,000. According to the First finance and Economics, cerebrovascular epidemiology data show that globally, the lifetime risk of stroke in people over the age of 25 is 24.9%, and the figure in China is close to 40%, that is, 40% of people are likely to have a stroke from the age of 25.

According to the announcement issued by the National Health Commission, the regional differences of the fluid characteristics of stroke are as follows: developing countries are greater than developed countries; The differences in the population are as follows: morbidity and mortality increase with the increase of age, the prevalence of stroke in males is generally higher than that in females in all countries, and the incidence of stroke in different races in the same region is significantly different. The disease factors associated with stroke include high blood pressure, heart disease, diabetes, transient ischemic attack, etc., poor lifestyle including smoking, excessive alcohol consumption, etc.

The purpose of this paper is to study the factors that have a great impact on stroke disease, and to achieve early intervention of the disease and reduce the prevalence rate through scientific methods to assist diagnosis. Combining the medical field

and machine learning can expand the application range of algorithms, enrich the theoretical basis, and provide new ideas for solving medical problems.

2.Research status at home and abroad

In 2002, Lumley[7] and other researchers conducted a 6.3 year follow-up of 5,711 residents over 65 years old with no history of stroke, constructed a sex-specific prediction equation using the characteristic variables most associated with stroke, and conducted experiments on a Web-based interactive platform to develop a model for predicting stroke in elderly Americans. In 2015, Xu Xiao et al. [1] established a fuzzy comprehensive evaluation model for stroke by quantifying fuzzy data based on various signs and factors of stroke and using genetic algorithm. In 2015, Manuel D. G. [8] and other researchers conducted a cohort survey of 82,259 Ontario residents since 2001, recording 3,236 stroke events. They constructed an index to comprehensively assess the impact of health habits and stress on stroke risk, which individuals can use for stroke risk assessment. In 2017, Lai Xinxing [2] standardized clinical data based on 547 cases of clinical trials, screened out independent risk factors for early neurological deterioration, and analyzed the relationship between DWI-ASPECTS scores and early neurological deterioration and its predictive ability. The correlation between neuroimaging and early neurological deterioration was analyzed by brain topography. In 2019, Wu Juhua [3] et al. identified 12 risk factors and built a neural network model for stroke risk prediction, and found 6 most important factors, such as total cholesterol and low density lipoprotein, with a prediction accuracy of 97.10%. In 2020, Luo Yishu [4] et al. proposed a multi-feature combined diagnosis model based on LSTM for clinical auxiliary diagnosis of ischemic stroke. The overall performance of the model reached 84%, which could provide a reference for doctors in differential diagnosis. In 2021, Hou Yumei [5] et al. built a prediction model of stroke incidence through data mining and Logistic regression model, enabling patients to self-monitor the risk of stroke occurrence and improving the convenience of stroke prevention. In 2023, Yang Huijie [6] et al. recorded the basic information of the enrolled patients, collected fasting blood samples to measure CRP, AIb and other indicators, and calculated CAR. Cervical DSA imaging was used to measure carotid artery stenosis, and the degree of stenosis on the most severe side was used as the assessment basis to study the relationship between the ratio of serum C-reactive protein and albumin and carotid artery stenosis in patients with acute cerebral infarction.

3.Data analysis

3.1 Data collection

There are 11 variables in the dataset, totaling 40,911 rows of data. The specific fields and meanings are shown in Table 1.

Table 1：Description of variables

variable	implication	value
sex	sex	0= male, 1= female
age	age	R
hypertension	hypertension	0= None, 1= Yes
heart_disease	Heart disease	0= None, 1= Yes
ever_married	Marital status	0= “No”, 1= “Yes”
work_type	Type of work	0= “Child”, 1= “government work”, 2= “never work”, 3= “private”, 4= “self-employed”
Residence_type	Residential type	0= “Rural”, 1= “urban”
avg_glucose_level	Residential type	R
bmi	Body mass index	R
smoking_status	Whether you smoke or not	0= “No”, 1= “Yes”
stroke	Have you had a stroke	0= “No”, 1= “Yes”

3.2 Data preprocessing

3.2.1 Missing value

Using the isnan() command in matlab, find the missing value. Three empty values are detected in the “sex” column, which

account for a very small proportion of the total data, less than 0.01%. Therefore, the disp(sum()) command is used to delete the value directly.

3.2.2 Outliers

First, descriptive statistics are performed on the data, and the total number, average value, standard deviation, minimum value, first quartile, second quartile, third quartile and maximum value of each variable are output. For the “age” variable, the minimum value is -9, which is obviously not common sense. So delete the rows where the “age” variable is less than or equal to 0. A total of 81 lines were deleted. It can be seen from the data that avg_glucose_level (blood sugar level) is high when it is 100-125 mg/dL, and bmi (body mass index) is overweight when it is 25-29.9. Descriptive statistics show that, The average value of avg_glucose_level is about 122.1, and the average value of bmi is about 30.4, which is not within the normal range. It is speculated that there are outliers. Draw a box plot to detect whether there are outliers. Draw box plots for the continuity variables “age”, “avg_glucose_level” and “bmi” respectively. There are 922 abnormal data in total, accounting for 2.25% of the original data set, which is relatively large. Consider using other methods to determine outliers.

Figure 1：age box diagram

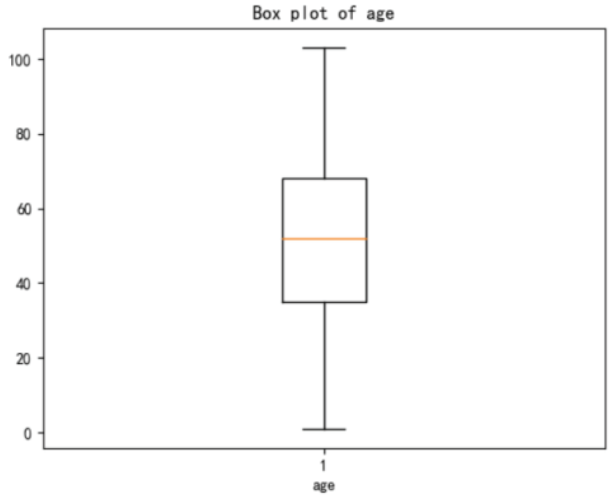


Figure 2 :bmi box plot

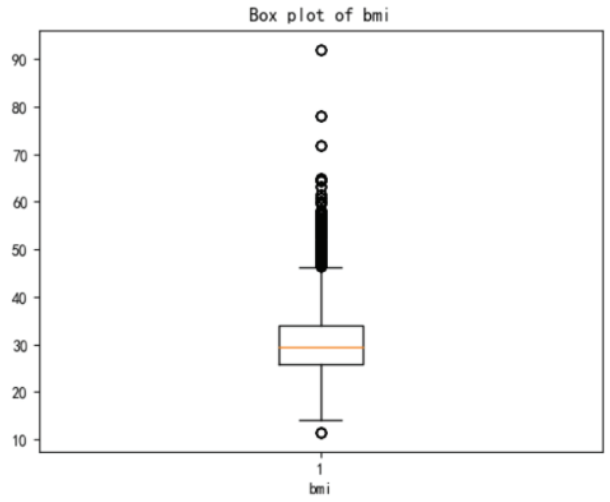
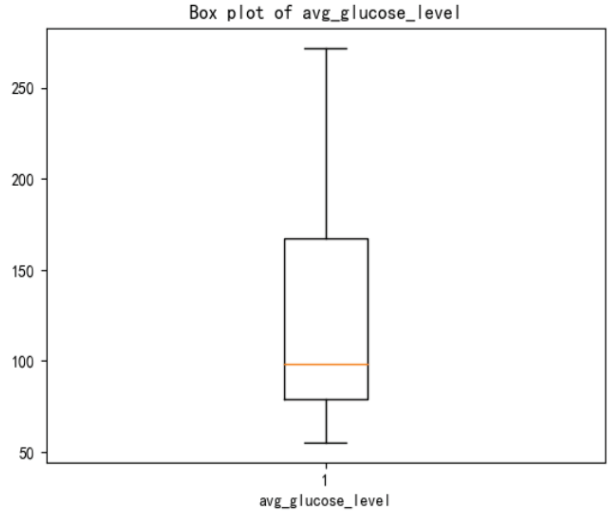


Figure 3:avg_glucose_level box diagram



Using the Laida criterion (3σ criterion) to find outliers, a total of 496 outliers were found.

For data points identified as outliers, further analysis of the source and possible causes. If the outliers are caused by actual physiological changes or disease conditions, retain these outliers and pay special attention to the impact of this part of the data in the analysis. In the case of outliers caused by measurement errors, the data combined with medical expertise should be cleaned to ensure that the handling of outliers is in line with the actual situation and will not cause loss or distortion of information. According to the relevant data of the National Population Health Sciences Data Center and the relevant data of

the interview of the experts of the Cardio-Cerebrovascular disease network, 391 outliers were deleted.

3.3 Data visualization

3.3.1 For discrete variables

In order to clearly show the relationship between each discrete variable and stroke, this paper visually shows the distribution and trend of data through cross-contingency tables, which can better identify the correlation and mutual influence between different variables.

Table 2:Cross contingency table of partial discrete variables and stroke

variable	value	stroke		total
		0	1	
sex	0	7870(43.662%)	10155(56.338%)	18025
	1	12286(54.824%)	10124(45.176%)	22410
total		20156	20279	40435
hypertension	0	18026(56.682%)	13776(43.318%)	31802
	1	2130(24.673%)	6503(75.327%)	8633
total		20156	20279	40435
heart_disease	0	19079(54.156%)	16151(45.844%)	35230
	1	1077(20.692%)	4128(79.308%)	5205
total		20156	20279	40435

According to Table 2, the probability of stroke is much higher in people with hypertension and diabetes than in people without stroke. The probability of stroke is 33.464% higher in people with diabetes and 31.919% higher in people with hypertension. The correlation between hypertension and heart_disease and stroke was high and hypertension and heart_ disease were judged. For the ever_married variable, the probability of having a stroke without being married was lower than that of being married. For the work_type variable, children and government workers have a very low probability of stroke, while the other types of work have high stress, which may increase the probability of stroke. For sex, Residence_type and smoking_status, the proportion of stroke and non-stroke is very balanced, and the impact of the four variables on stroke cannot be directly judged.

3.3.2 For continuous variables

The density curves of three variables, age,avg_glucose_level and bmi, and stroke, can be drawn. It can be found that age and bmi are normally distributed, which accords with the real data.

As can be seen from Figure 4, when the value of age variable is greater than 70, the proportion of stroke patients is significantly higher than that of non-stroke patients.

Normal blood glucose usually refers to the fasting blood glucose concentration in milligrams per deciliter (mg/dL). The generally accepted normal range of blood sugar is postprandial blood sugar (after two hours) : less than 140 mg/dL. Figure 5 shows that when the value of avg_glucose_level variable is greater than 170, the proportion of stroke patients is significantly higher than that of non-stroke patients. This suggests that hyperglycemia can increase the incidence of stroke and is an independent risk factor for stroke.

The normal BMI for adults is between 18.5 and 23.9, and if a BMI below 18.5 is considered underweight, a BMI of 24 to 27 is overweight, and a BMI of 28 to 32 is obese. If your BMI is over 32, you are considered obese. Figure 6 shows that when the bmi variable is between 24 and 30, the proportion of people who have had a stroke is significantly higher than that of people who have not had a stroke. This suggests that the higher the obesity, the higher the incidence of stroke.

Figure 4: Curve of stroke value frequency and age density

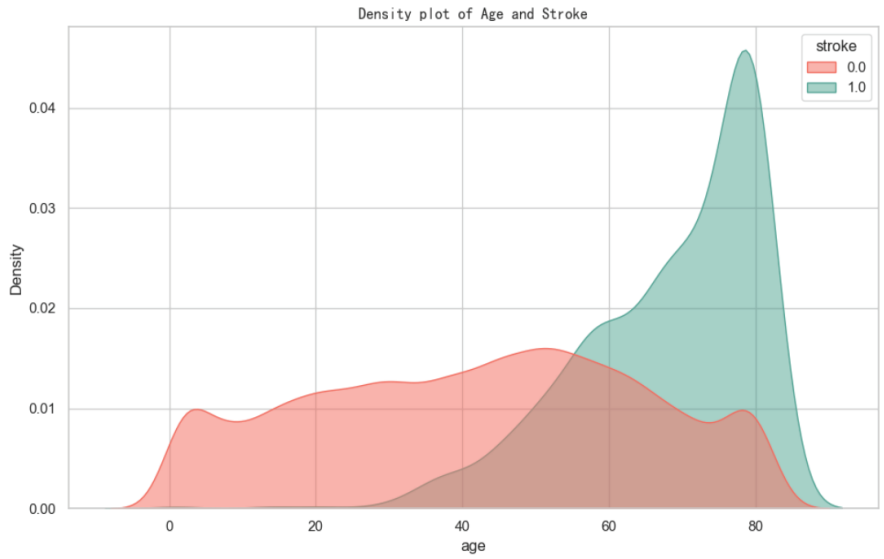
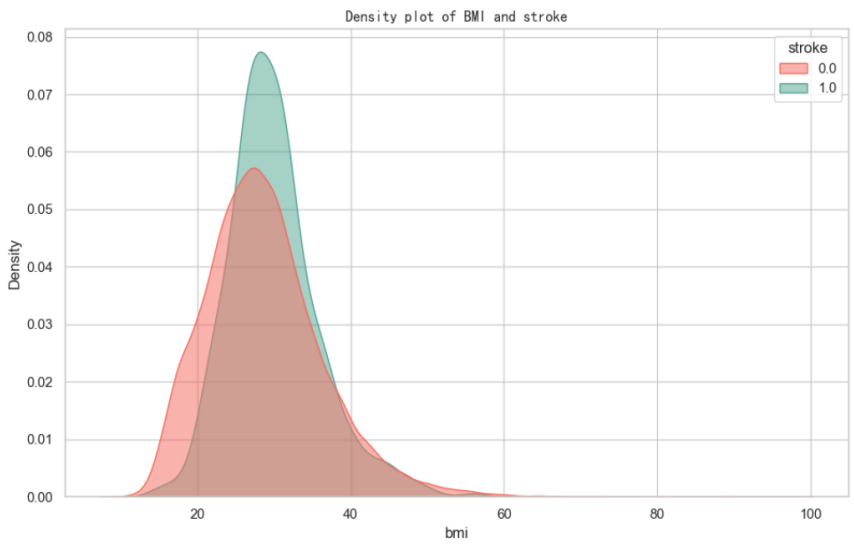


Figure 5: Graph of the value frequency of stroke and the density of avg_glucose_level

Figure 6: Curves of stroke value frequency and bmi density



3.4 Correlation analysis

3.4.1 For discrete variables

The Pearson Chi-square test enables statistical significance levels and effect sizes, making it easier to comprehensively evaluate the relationship between variables. Pearson Chi-square test is suitable for data where the sample size is large enough and both categorical variables are discrete variables. Therefore, Pearson Chi-square test can be used for correlation analysis of discrete variables in this data.

The chi-square test results are as follows:

Table 3:Results of chi-square test between stroke and some categorical variables

variable	value	stroke		total	Inspection method	X^2	P
		0	1				
sex	0	7870	10155	18025	pearson Chi-square test	497	0.000***
	1	12286	10124	22410			
total		20156	20279	40435			
variable	value	stroke		total	Inspection method	X^2	P
		0	1				
hypertension	0	18026	13776	31802	pearson Chi-square test	2782	0.000***
	1	2130	6503	8633			
total		20156	20279	40435			
variable	value	stroke		total	Inspection method	X^2	P
		0	1				
heart_disease	0	19079	16151	35230	pearson Chi-square test	2031	0.000***
	1	1077	4128	5205			
total		20156	20279	40435			
Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively							

From the above analysis results, it can be seen that the correlation between stroke and Residence_type is not significant, and there is a highly significant correlation with other variables.

3.4.2 For continuous variables

This paper aims to study the correlation between three continuous variables and the stroke variable. As mutual information does not rely on the distribution assumption of data and has good robustness to outliers and noise data, it is adopted for the correlation analysis.

Table 4 Mutual Information Coefficient

variable	Stroke
age	0.017
avg_glucose_level	0.667
bmi	0.228

From Table 4, it can be seen that all three variables have a positive relationship with the stroke variable. The variable of avg_glucose_level has the highest correlation with the stroke variable. The bmi variable also has an impact on stroke, while the age variable has the least impact.

Bivariate analysis can reveal the direction of the association (positive correlation, negative correlation, or no correlation) and

the degree of association (correlation coefficient) between two continuous variables. In this paper, plots were made for the pairwise relationships between the three variables of age, avg_glucose_level, bmi and stroke. As can be seen from Figures 7, 8 and 9, when the age is above 60 years old, regardless of the blood glucose level and the BMI index, the probability of stroke is very high. When the blood glucose level is higher than 200, regardless of the age and the BMI index, the probability of stroke is also relatively high.

Figure 7 Scatter Plot of Age and Blood Glucose Level

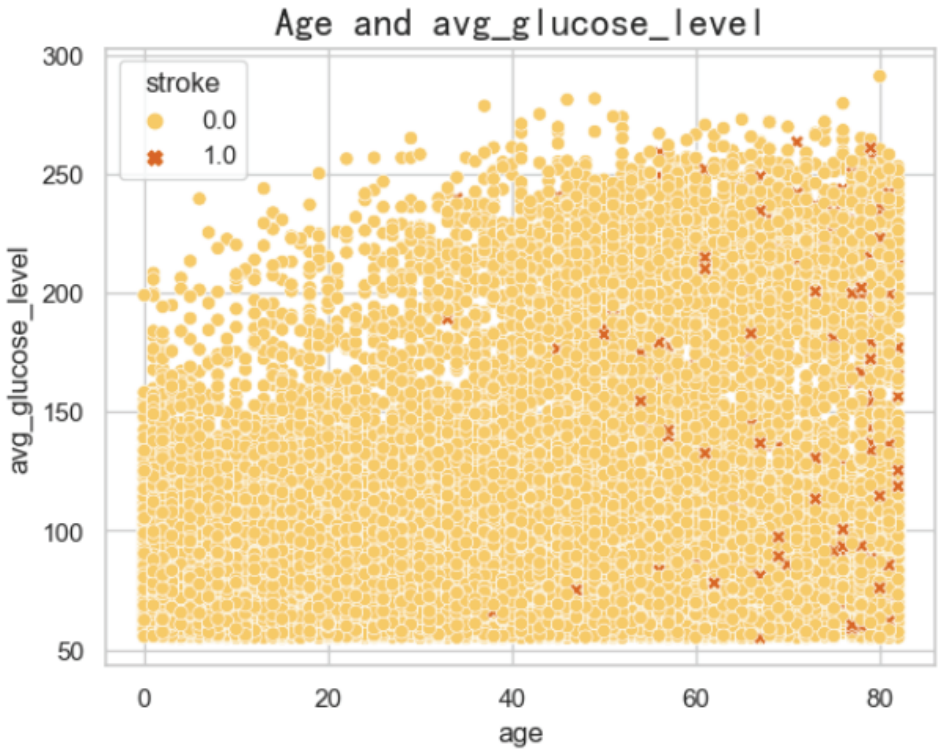


Figure 8 Scatter Plot of Age and BMI Index

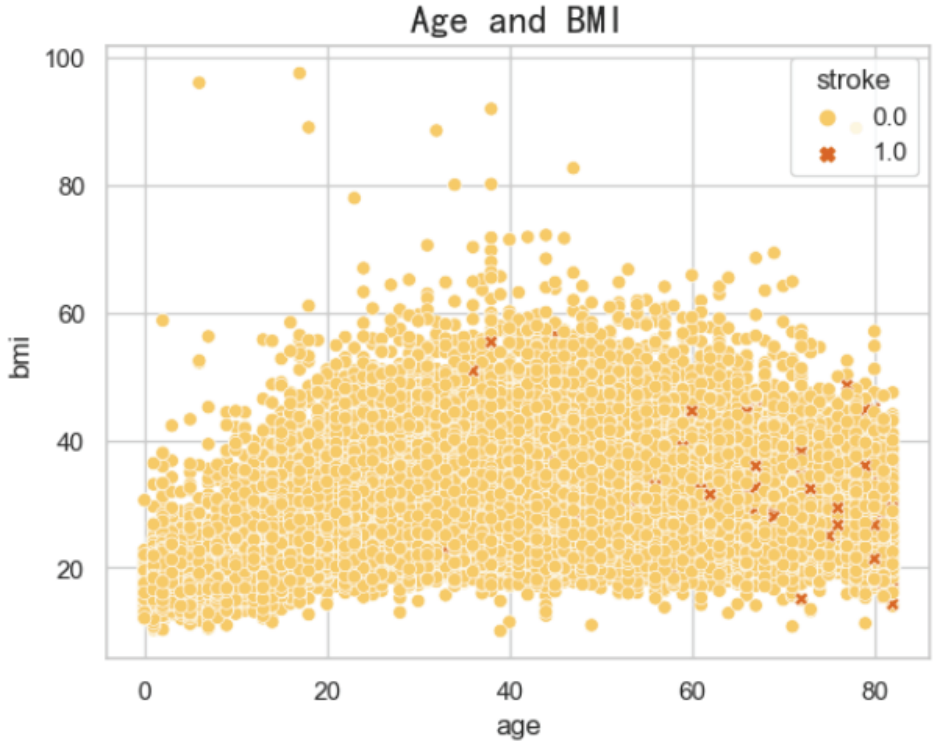
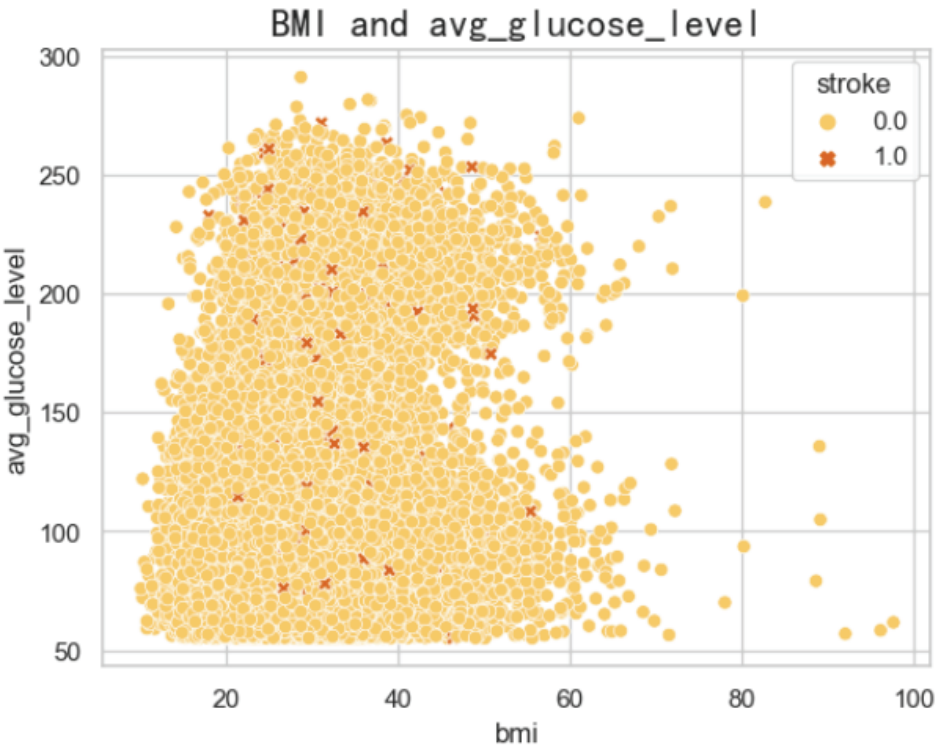


Figure 9 Scatter Plot of BMI Index and Blood Glucose Level



4.Stroke Risk Prediction

4.1 Support Vector Machine

The support vector machine is a binary classification model that can handle high-dimensional data and has a relatively high prediction accuracy. In stroke prediction, data of individuals with strokes and those without strokes can be used as samples, and classification can be carried out through the SVM algorithm to predict whether a person will have a stroke. According to the confusion matrix, the number of True Positives (TP) is 3223, the number of False Negatives (FN) is 796, the number of False Positives (FP) is 890, and the number of True Negatives (TN) is 3178.

Table 5 Evaluation Indicators of the Support Vector Machine

Support Vector Machine Indicators	Values
Accuracy	0.792
Precision	0.784
Recall	0.802

From the results in Table 5, it can be seen that although the accuracy and precision of the support vector machine are relatively low, and it fails to achieve an excellent classification effect.

4.2 Random Forest

Random forest is an ensemble learning algorithm. It is a classifier that contains numerous decision trees, and it can synthesize the prediction results of decision trees to improve the accuracy of the model. Random forest can effectively deal with overfitting and handle high-dimensional data. It has good robustness and interpretability and is suitable for classification and regression problems. According to the confusion matrix, the number of True Positives (TP) is 4018, the number of False Negatives (FN) is 1, the number of False Positives (FP) is 24, and the number of True Negatives (TN) is 4044.

Table 6 Evaluation Indicators of Random Forest

Random Forest Indicators	Values
Accuracy	0.994
Precision	0.997
Recall	1

From the results in Table 6, it can be seen that the accuracy and precision of the random forest are extremely high, exceeding 99%. The classification results obtained using the random forest are highly persuasive.

The random forest can output feature importance. Feature importance refers to the contribution degree of each feature to the prediction result of the model. Feature importance is calculated by measuring the degree of reduction in the Gini index brought by each feature when the model splits each node of the decision tree. The lower the Gini index, it indicates that the model has a higher degree of dependence on this feature, so the feature importance is also higher.

Table 7 Ranking of Feature Importance

Variable	Importance
age	0.277742
avg_glucose_level	0.222129
bmi	0.059261
work_type	0.053656
hypertension	0.037157
Residence_type	0.036793
smoking_status	0.034687
heart_disease	0.032677
ever_married	0.029671
sex	0.007046

As can be seen from Table 7, among the continuous variables, age, avg_glucose_level, and bmi contribute the most to stroke. Among the discrete variables, work_type contributes the most to stroke, which is consistent with the results of the chi-square test. In contrast, sex has the least impact on stroke. 4.3 BP Neural Network

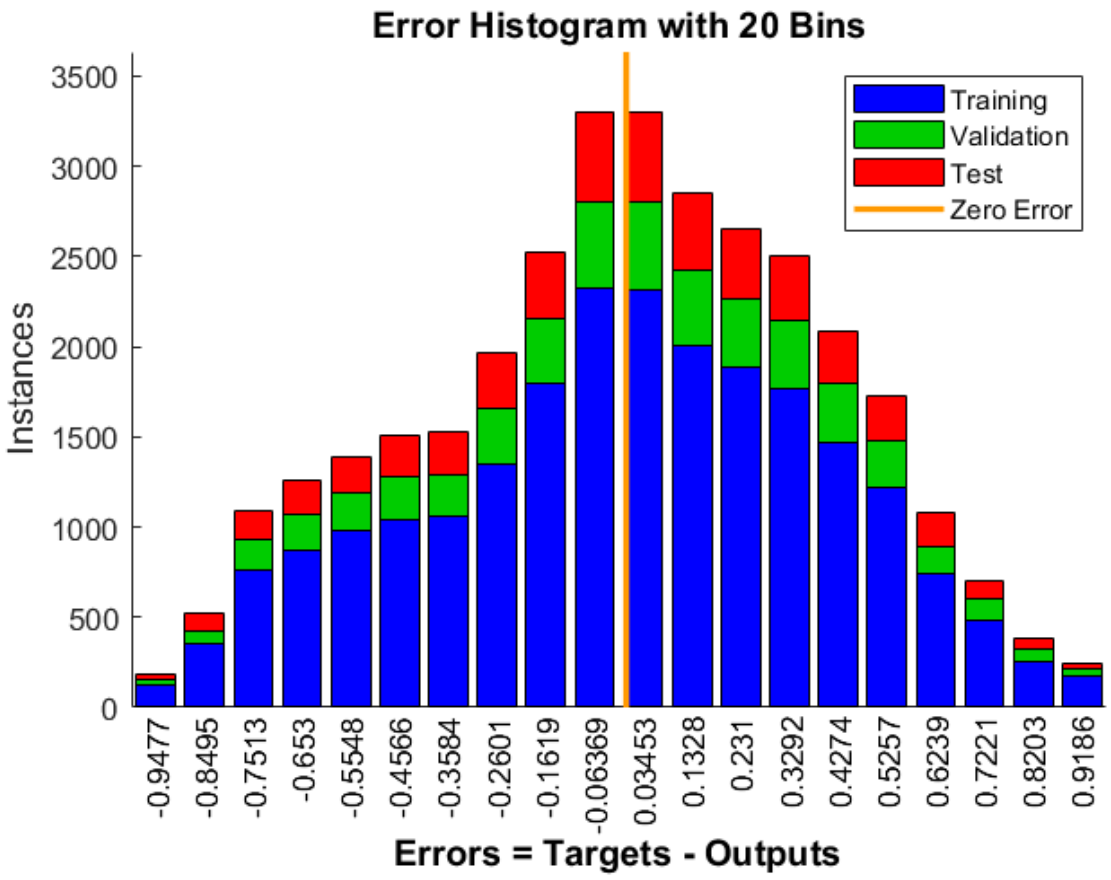
The main components of a BP neural network include the input layer, hidden layer, and output layer. The input layer is responsible for receiving the original data, the hidden layer is responsible for feature extraction and data transformation, and the output layer is responsible for generating the final prediction results. During the training process, the BP algorithm calculates the error between the prediction results and the actual targets, and then propagates the error backward from the output layer to the input layer, adjusting the network weights and biases layer by layer to minimize the error. By adjusting the training function and transfer function, it can accurately predict whether a person has a stroke. According to the confusion matrix, for the BP neural network as a whole, the number of True Positives (TP) is 12300, the number of False Negatives (FN) is 4087, the number of False Positives (FP) is 3457, and the number of True Negatives (TN) is 12953.

Before optimizing the neural network with the heuristic algorithm, both the accuracy and precision of the model are lower than 0.8, resulting in a poor prediction effect.

Table 8 Evaluation Indicators of the Neural Network

Neural Network Indicators	Values
Accuracy	0.760
Precision	0.520
Recall	0.789

Figure 15 Error Histogram



The closer the error is to 0, the better. As can be seen from Figure 15, a large portion of the data falls outside the range of -0.4 and 0.4. It can also be observed that the prediction effect of the neural network before optimization is poor.

4.4 Neural Network Optimized by Genetic Algorithm

The genetic algorithm is an optimization method that simulates the biological evolution process. The solutions to the problem are represented as chromosomes, and an initial population of solutions is randomly generated. By evaluating the fitness of the chromosomes, the better chromosomes are selected for reproduction, and operations such as crossover and mutation are carried out to generate new chromosomes. Through iterative evolution, the chromosome with the highest fitness in the population is finally output as the optimal solution.

According to the confusion matrix, the number of True Positives (TP) is 198, the number of False Negatives (FN) is 0, the number of False Positives (FP) is 0, and the number of True Negatives (TN) is 7984.

Table 9 Evaluation Indicators of the Neural Network Optimized by the Genetic Algorithm

Indicators of the Neural Network Optimized by the Genetic Algorithm	Values
Accuracy	0.933
Precision	1.00
Recall	0.933

It can be seen from the results in Table 9 that the accuracy, precision and recall rate of the neural network optimized by the genetic algorithm are extremely high, exceeding 90%. The neural network optimized by the genetic algorithm has achieved excellent results in the classification task.

4.5 Comparison of Model Performance

Table 10 Comparison of Model Performance

	Precision	Advantages	Disadvantages
Support Vector Machine(SVM)	0.777	Insensitive to outliers.	The training speed is relatively slow.
Random Forest	0.998	It performs well in handling the interactions between features and high - dimensional datasets, and can obtain the importance of features.	It may overfit.
BP (Back Propagation) Neural Network	0.520	The hidden layer can learn new features that contribute to classification.	It requires a large amount of parameter adjustment and experimentation.
Neural Network Optimized by Genetic Algorithm	1.00	Global search finds better solutions for neural networks.	The computational cost is relatively high.

According to the data in Table 10, the precision rates of both the Support Vector Machine and the BP Neural Network fail to reach 0.8, indicating relatively low classification accuracy. The Random Forest and the Neural Network optimized by the Genetic Algorithm perform excellently, with their precision rates both exceeding 0.99, demonstrating outstanding performance in classification tasks. However, the Random Forest has the risk of overfitting, while the Neural Network optimized by the Genetic Algorithm can find better solutions through global search, providing potential advantages for improving classification accuracy. By comparing factors such as the performance, interpretability, and computational cost of different models, the Neural Network model optimized by the Genetic Algorithm is finally adopted, achieving the best classification effect and prediction ability.

5.Conclusions and Recommendations

5.1 Conclusions

1.There is an obvious correlation between age and stroke. Especially in the age group of 90 years old and above, the incidence rate of stroke increases significantly, accounting for about 90% of the population in this age group. This indicates that the risk of stroke increases significantly with the growth of age.

2.Blood glucose level is closely related to stroke. When the average blood glucose level is higher than 170, the proportion of stroke patients is significantly higher than that of the non-stroke population. High blood glucose level may be a manifestation of diabetes, and the risk of stroke among diabetic patients is much higher than that among people with normal blood glucose levels. When the blood glucose concentration exceeds 150, regardless of the BMI value, the risk of stroke increases significantly.

3.The BMI index is also associated with stroke. When the BMI index is greater than 21, the number of stroke patients gradually increases. Especially when the BMI exceeds 27, the proportion of stroke cases is higher than that of non-stroke cases.

4.The comprehensive results of the feature importance ranking of the random forest and the chi-square test show that age, hypertension, and heart disease have a very significant impact on stroke, and there is a close correlation between these factors and stroke.

5.2 Recommendations

1.Regularly monitor blood pressure, blood glucose, and BMI index, and keep them within the normal range. Control your diet, avoid high-sugar and high-fat foods, and maintain an appropriate weight. Eat more vegetables, fruits, whole grains, and low-fat dairy products, and reduce the intake of saturated fats and trans fats.

2.Actively prevent hypertension and heart disease by strengthening physical exercise, quitting smoking and limiting alcohol consumption, maintaining mental health, etc. Seek medical advice in a timely manner and receive guidance and treatment from professional doctors.

3.Maintain an appropriate weight. Use diet and exercise to maintain a normal weight and avoid obesity.

4.For patients with heart disease, follow the doctor's advice for drug treatment and lifestyle adjustments. Pay attention to heart symptoms such as chest pain, shortness of breath, dizziness, etc., and seek medical attention promptly.

5.Pay attention to the adjustment of lifestyle. Keep a regular work and rest schedule, avoid excessive fatigue and stress, and maintain a peaceful mindset.

6.Strengthen health education, improve the awareness of stroke risks, cultivate healthy living habits and behaviors, and enhance the awareness of self-care.

Funding

no

Conflict of Interests

The author(s)declare(s) that there is no conflict of interest regarding the publication of this paper.

References

[1] Xu Xiao, Zhang Lei, Wang Zhen. Brain Stroke Risk Prediction System Based on Fuzzy Comprehensive Evaluation Method[J]. Computer Simulation, 2015, 32(07): 344-347 + 360.

[2] Lai Xinxing. Study on Risk Factors and Prediction Model of Early Neurological Deterioration in Acute Ischemic Stroke[D]. Beijing University of Chinese Medicine, 2017.

[3] Wu Juhua, Zhang Shuo, Tao Lei, et al. Research on Stroke Risk Prediction Model Based on Neural Network[J]. Data Analysis and Knowledge Discovery, 2019, 3(12): 70-75.

[4] Luo Yishu, Shao Yuanyuan, Chen Dehua. Ischemic Stroke Diagnosis Model Based on the Combination of Multiple Features of LSTM[J]. Intelligent Computer and Its Applications, 2020, 10(10): 74-79.

[5] Hou Yumei, Zeng Hui, Zhang Chenyang, et al. Prediction of the Risk of Ischemic Stroke Based on Data Mining[J]. Chinese Journal of Gerontology, 2021, 41(01): 177-181.

[6] Yang Huijie, Dong Jiankai, Hu Quanzhong. Study on the Relationship between the Ratio of Serum C-reactive Protein to Albumin and Carotid Artery Stenosis in Patients with Acute Cerebral Infarction[J]. International Journal of Laboratory Medicine, 2023, 44(08): 956-960.

[7] Thomas, Lumley, and, et al.A stroke prediction score in the elderly: validation and Web-based application[J].Journal of Clinical Epidemiology, 2002.DOI:10.1016/S0895-4356(01)00434-6.

[8] Manuel D G, Meltem T, Richard P, et al.Predicting Stroke Risk Based on Health Behaviours: Development of the Stroke Population Risk Tool (SPoRT)[J].Plos One, 2015, 10(12):e0143342.DOI:10.1371/journal.pone.0143342.