WEP Warwick Evans Publishing

Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

Risk Assessment of hematoma expansion in patients with hemorrhagic Stroke based on Stochastic Forest Prediction Model

Yuwei Xiong *, Xiaoyi Feng, Feiran Gao

School of Mathematics and Statistics, Yunnan University, Kunming, China, 650504

* Corresponding Author Email: xiongyuwei605@163.com

Abstract. This paper discusses the prediction model of hematoma expansion probability in patients with hemorrhagic stroke. Hemorrhagic stroke is an acute and critical neurological disease. High mortality and disability make it a major challenge in the field of public health. Hematoma dilatation and peri-hematoma edema are important factors leading to secondary brain injury. therefore, accurate prediction of the risk of hematoma expansion in patients with hemorrhagic stroke has important guiding significance for clinical diagnosis and treatment. First of all, this study carries on the matching integration and abnormal processing of the patient data in order to improve the data quality. Then, a classification model based on support vector machine (SVM) and random forest is constructed to predict the probability of hematoma expansion. Through the comparative analysis of the classification accuracy, it is found that the prediction effect of the classification model based on random forest is better than that based on SVM. This result provides clinicians with a more accurate prediction tool, which is helpful to identify high-risk patients early, reduce the incidence of secondary brain injury, and improve the quality of life and prognosis of patients. This study provides a new idea for the clinical diagnosis, treatment and research direction of hemorrhagic stroke, and has important theoretical and practical value.

Keywords: Hemorrhagic stroke; Hematoma expansion; Predictive model; Support vector machine (SVM); Random Forest.

1. Introduction

Hemorrhagic stroke is a serious neurological disease, which refers to the bleeding caused by vascular rupture in the brain parenchyma, and is one of the most fatal types of strokes. Its incidence accounts for 10-15% of all stroke cases. The pathogenesis of hemorrhagic stroke is due to the rupture of arteries or veins in the brain, causing blood to enter the brain tissue. This can cause stress and damage to surrounding brain cells because the blood contains hemoglobin, which can cause inflammation and oxidative stress in brain tissue. Hemorrhagic stroke may lead to severe neurological deficits, including paralysis, speech disorders, cognitive impairment and loss of the ability to live independently, which not only seriously affect the quality of life of patients, but also cause varying degrees of burden on society [1, 2]. Studies have shown that about 1/3 of the patients with hemorrhagic stroke will have hematoma enlargement shortly after the onset of symptoms, and the hematoma enlargement of hemorrhagic stroke is one of the main reasons for the deterioration of clinical condition and poor prognosis. An increase in hematoma volume exceeding 33% of the basal hematoma volume or more than 6 mL is defined as hematoma enlargement. The prognosis of enlarged hematoma depends on a number of factors, including the size, location, cause and treatment of the hematoma. Early diagnosis and treatment can improve the prognosis of patients, but severe enlargement of hematoma may lead to long-term neurological damage and serious complications. In addition, peri-hematoma edema is also an important factor leading to secondary brain injury. Perihematomal edema refers to an inflammatory reaction in the brain tissue around hemorrhagic lesions in the brain.

Edema around hematoma can cause temporary ischemia around blood clot, destruction of blood-brain barrier, decrease of metabolic activity of brain tissue, and serious nerve injury [3, 4]. Peri-hematoma edema is closely related to the prognosis of patients with hemorrhagic stroke.

Therefore, it is of great clinical significance to identify hemorrhagic stroke patients early and effectively, and to predict the risk of hematoma enlargement and perihematomal edema after intracerebral hemorrhage, so as to implement anti-expansion and intervention treatment.

2. Research methods.

First of all, the data is preprocessed. The main purpose of data preprocessing is to match the personal information of each patient with the examination time, examination content and examination result data, and to find out whether the data has missing values, outliers, data dislocation and so on.

After preprocessing the data, we judged whether the patients had hematoma expansion according to the changes of hematoma volume, and provided training data for the follow-up questions. Then, we analyzed the factors related to the patients' personal history, disease history, treatment methods and image examination results, trained SVM support vector machine model and random forest model, and then used the model to predict the probability of hematoma expansion in all patients.

Finally, after comparing the prediction accuracy of the two models, the model with higher prediction accuracy is selected to predict the probability of hematoma expansion in patients.

3. Model establishment and solution.

3.1. Data preprocessing.

To determine whether hematoma dilatation occurred in patients with hemorrhagic stroke within 48 hours after onset, and if so, the time of hematoma dilatation should be recorded at the same time.

First of all, we carried out data pre-processing to extract the required data, such as the serial number of the first image examination of the first 100 patients, the time point of the first examination, the time from onset to the first image, and so on. No data missing, data duplication and other problems were found.

The time interval from the onset of the first 100 patients to the first examination and multiple follow-up examinations, and the absolute volume and relative changes of hematoma compared with the first examination were calculated.

The formula of absolute volume change is $V_i - V_0$, and the formula of relative volume change is $\frac{V_i - V_0}{V_0}$. If the relative volume change of one follow-up is "- 1", it indicates that the follow-up has not been carried out, and replace it with "0".

The results of data preprocessing are shown in Table 1.

Table 1. partial results of data preprocessing are shown

Patient number	Follow-up 1 time difference	Follow-up 1 changes of absolute volume of hematoma	Follow-up 1 changes of relative volume of hematoma	Follow-up 2 time difference	Follow-up 2 changes in absolute volume of hematoma	Follow-up 2 changes of relative volume of hematoma
sub001	8.28	5.188	7.44%	132.12	1.238	1.78%
sub002	14.92	4.771	10.04%	69.22	0.248	0.52%
sub003	9.53	19.646	22.74%	39.60	16.867	19.52%
sub004	16.99	-5.621	2.35%	83.86	-28.876	-63.47%
sub005	26.48	9.64	64.99%	97.96	10.645	71.77%
sub006	47.88	-23.037	3.46%	121.64	-97.042	-56.71%
sub007	44.09	-0.729	-2.48%	117.14	-2.101	-7.14%
sub008	10.26	5.878	21.28%	98.26	9.239	33.45%
sub009	40.06	15.272	36.94%	66.16	12.572	30.41%
sub010	15.00	-9.006	-24.29%	39.34	-21.666	-58.44%

3.2. Identification and time screening of hematoma dilatation.

According to the relevant definition of "hematoma dilatation", the patients with hematoma absolute volume increase ≥ 6ml or hematoma relative volume increase ≥ 33% were selected according to the data of each follow-up examination which occurred within 48 hours, and recorded as "hematoma expansion event".

Taking follow-up 1 as an example, first, according to the time interval of follow-up 1 obtained by data preprocessing, the follow-up events with an interval of less than 48 hours were screened, and then the changes of hematoma absolute volume increase ≥ 6ml or relative volume increase ≥ 33% were screened, and the eligible patients were marked yellow and recorded.

And so on, the other follow-up procedures are the same.

3.3. Time record of hematoma dilatation.

For patients with hematoma dilatation, record the difference in follow-up time when hematoma dilatation was found, that is, the interval between the follow-up time and the onset time of hematoma dilatation.

Here, we also have a deep thinking, during the follow-up, we found that the hematoma dilated, the follow-up time is not necessarily the real time of hematoma expansion, may have occurred earlier than the follow-up time, only found during the follow-up.

In order to further explore the real time of hematoma dilatation, we analyzed and verified it by linear regression.

Analysis and verification based on linear regression: input the integrated data into python from the known information, because only the interval between the first two examinations and the onset time is within 48 hours, so only need to extract the time of onset, the time of the first and second follow-up examination and the corresponding hematoma volume data.

Taking the time as the horizontal axis and the absolute volume increase data of the first inspection as the vertical axis, the following scatter plot can be drawn, as shown in figure 1.

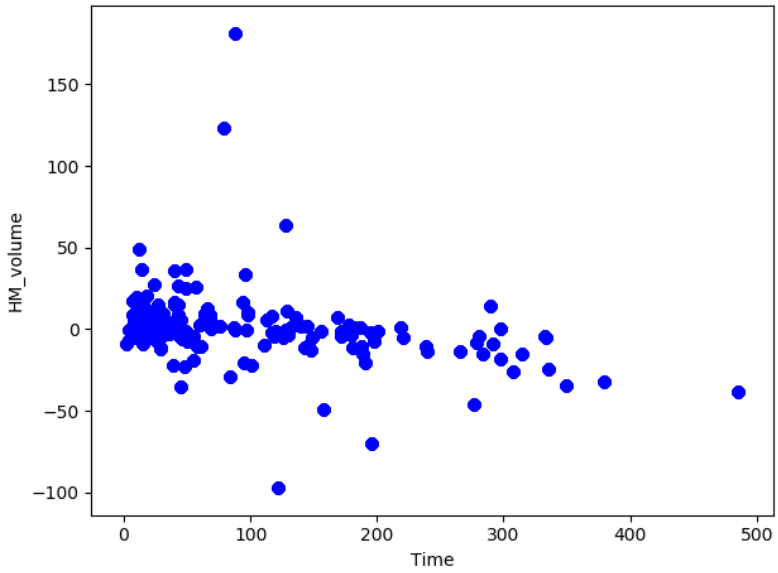

Figure 1. scatter plot of raw data.

Because the coordinate axis is set with the onset time as the origin, and the first follow-up examination is mostly during the hospitalization of patients, the interval is relatively short, so the left side of the data is more intensive.

As can be seen from the above picture, the distribution of the data has a certain trend and is more concentrated, so it is fitted by linear regression.

The time of the two follow-up examinations and the corresponding absolute changes of hematoma volume were combined as fitting data to construct a univariate linear regression. The results are shown in Figure 2 below.

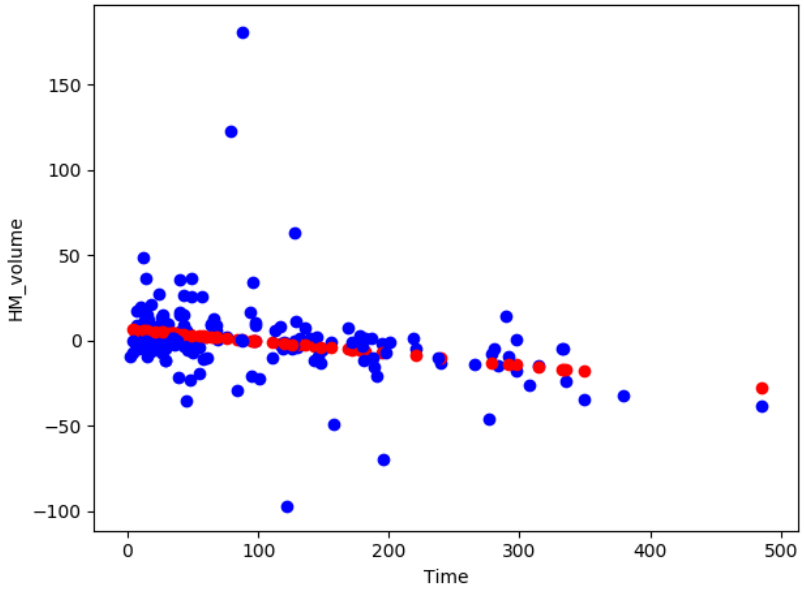

Figure 2. Linear fitting equation.

It can be seen that except for a few extreme ends, the fitting degree of the equation is good, which is the same as that of most patients, so the data can be analyzed by linear regression.

Each patient was modeled by linear regression, and their linear regression equations of hematoma volume were obtained respectively.

Through the linear regression equation, we can further explore the time point of hematoma dilatation. We make y meet the conditions of hematoma dilatation and solve the x value, which is not much different from the time recorded before. It is further verified that the time of occurrence of hematoma dilatation recorded by us is correct.

3.4. Predicting the probability of hematoma dilatation.

Construction of classification prediction model: based on the above data preparation work, we plan to use the image examination data of the first 100 patients to train the optimal classification prediction model, and then use this model to predict the probability of hematoma expansion in all patients.

Therefore, we divide the data integration table into the table used in the training model and the table used to predict probability.

3.5. Prediction probability of constructing classification model based on SVM.

3.5.1. Brief introduction of support Vector Machine (SVM) algorithm.

Support Vector Machine (SVM) is a classification method proposed by Vapnik et al. It was originally designed for binary classification. It is a supervised machine learning method, which is often used in target recognition, data mining and other fields.

By solving the quadratic programming optimization problem, the support vector machine classifier fits the widest distance between the two categories, so that the distance between the two dotted lines on the way is maximum.

It constructs a separate hyperplane to perform the calculation operation and classifies the input data. The hyperplane divides the data set into two or more classes based on the label class, and the points of different shapes on the left and right sides of the hyperplane represent different types.

The decision plane is a multi-dimensional hyperplane, which divides the data set into different classes.

In the linear classification problem, two parallel hyperplanes are found to optimally separate categories, as shown in figure 3.

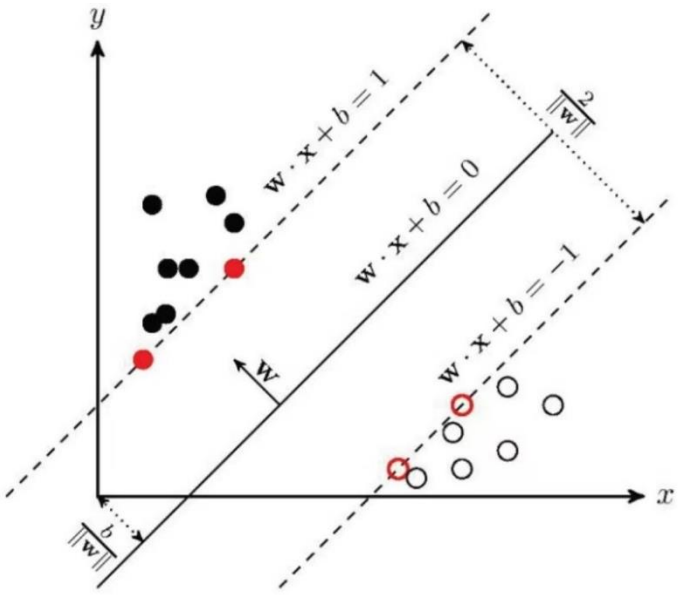

Figure 3. Optimal classification hyperplane

Given two samples x_i, y_i, i=l, ···m, y_i E {-1,1}, then the linear equation of the hyperplane is:

$$w \cdot x + b = 0 \tag{1}$$

Divide the two simultaneous hyperplane sums parallel to H by the linear equation above:

$$\begin{cases} w \cdot x + b = 1 \\ w \cdot x + b = -1 \end{cases} \tag{2}$$

The constraints on the hyperplane H are:

$$\begin{cases} w \cdot x_i + b \geq 1, y_i = 1 \\ w \cdot x_i + b \geq 1, y_i = -1 \end{cases} \tag{3}$$

According to the distance equation, the distance between hyperplanes H1 and H2 is obtained:

$$y = \frac{2}{\|w\|} \tag{4}$$

The objective of SVM is to find the dividing hyperplane with maximum interval, i.e., to find the parameters w and b that satisfy the constraints of (4) such that the classification interval is maximum. Establish the optimization objective:

$$\min = \frac{1}{2}\|w\|^2 \tag{5}$$

This is the basic model of a support vector machine. The objective function is a quadratic optimization problem, so the Lagrangian method is used to solve for the extremes using the following formula.

$$L(w, b, a_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i(y_i(w \cdot x + b) - 1) \tag{6}$$

Convert it to a dyadic problem:

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j a_i a_j (x_i \cdot x_j) - \sum_{i=1}^{n} a_i \tag{7}$$

$$s.t. \sum_{i=1}^{m} a_i y_i = 0 \tag{8}$$

$$0 \leq a_i \leq C, i = 1, \cdots\cdots n \tag{9}$$

Solve the above optimization problem solution as a_i^*, when $0 < a_i^* < C$, a_i^* is a standard support vector, when the KKT condition is:

$$a_i^*[y_i(w^* \cdot x_i) + b^* - 1] = 0 \tag{10}$$

Solve for b^* as:

$$b^* = y_i - w^* \cdot x_i = y_i - \sum_{i=1}^{n} y_i a_i^* (x_i \cdot x_j) \tag{11}$$

Get the decision function:

$$f(x) = \text{sign}(\sum_{i=1}^{n} y_i\, a_i^*(x_i \cdot x) + b^*) \tag{12}$$

3.5.2. Kernel function

In practical applications, most of the classification problems involved are nonlinear, and in order to solve nonlinear classification problems, the concept of kernel function is introduced into machine learning so as to map the sample data from a low-dimensional feature space to a high-dimensional feature space, which makes the nonlinear mapping more easily realized.

A nonlinear transformation is performed on the feature x. The new feature is denoted as $z = \varphi(x)$, and the corresponding quadratic optimization problem becomes:

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j a_i a_j\big(\varphi(x_i) \cdot \varphi(x_j)\big) - \sum_{i=1}^{n} a_i \tag{13}$$

$$s.t. \sum_{i=1}^{m} a_i y_i = 0 \tag{14}$$

$$0 \le a_i \le C, i = 1, \cdots\cdots n \tag{15}$$

The decision function obtained at this point is:

$$f(x) = \text{sign}(\sum_{i=1}^{n} y_i\, a_i^*(\varphi(x_i) \cdot \varphi(x)) + b^*) \tag{16}$$

Mapping the way the inner product $(x_i \cdot x_j)$ is computed in the low-dimensional feature space to the inner product $(\varphi(x_i) \cdot \varphi(x_j))$ in the high-dimensional feature space, the kernel function can be utilized to compute this inner product, denoted as:

$$K\big(x_i, x_j\big) = \Big(\varphi(x_i) \cdot \varphi\big(x_j\big)\Big) \tag{17}$$

The quadratic optimization problem at this point becomes:

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j a_i a_j K\big(x_i \cdot x_j\big) - \sum_{i=1}^{n} a_i \tag{18}$$

The decision function is:

$$f(x) = \text{sign}(\sum_{i=1}^{n} y_i\, a_i^* K(x_i \cdot x) + b^*) \tag{19}$$

There are several common kernel functions:

Linear kernel functions:

$$K(x, y) = (x \cdot y) \tag{20}$$

Polynomial kernel functions:

$$K(x, y) = \big((x + y) + 1\big)^d, d > 0 \tag{21}$$

Sigmoid kernel function:

$$K(x, y) = \tanh[a(x \cdot y) + b] \tag{22}$$

RBF kernel function:

$$K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{23}$$

3.5.3. Construct SVM classification model

Model construction idea: through the integrated dataset, the dataset is proportionally sliced into the training set and the test set, and the model is evaluated by the model evaluation indexes with different slicing ratios, such as accuracy, recall, precision, F1 index score, etc., so as to select the optimal model, which is used for predicting whether the hematoma dilatation event occurs in all the patients, and the probability of hematoma dilatation occurs is calculated.

The SVM classification prediction model was constructed using spsspro, and the model parameters were continuously adjusted to fit the optimal model, which was finally determined to be modeled with a training set of 0.8 and a test set of 0.2. The following outputs were obtained (Figure 4).

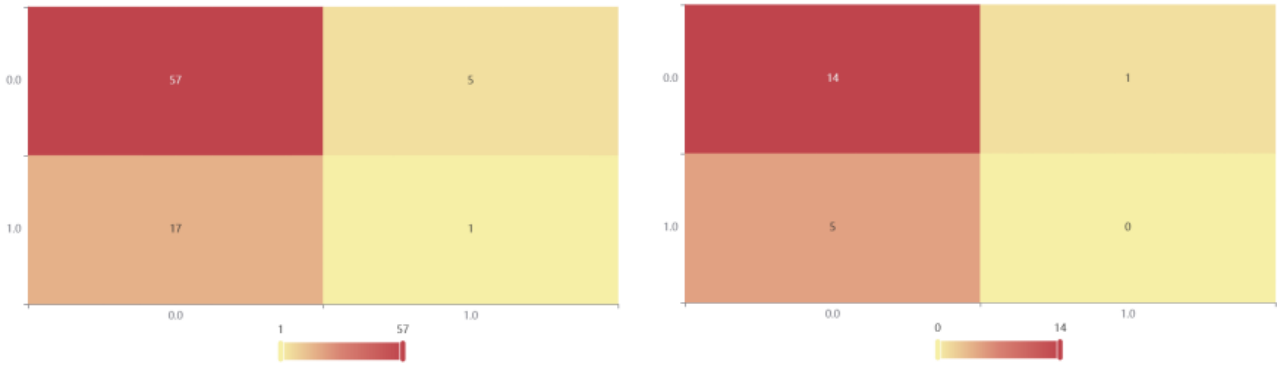

Figure 4. Heat map of confusion matrix for SVM classification model for training set (left) and test set (right)

Result analysis: the confusion matrix of the test set is shown in the form of a heat map, in which the number of correct judgments in the training set is 58 and the number of incorrect judgments is 22, and the number of correct judgments in the test set is 14 and the number of incorrect judgments is 6. The results of the evaluation of the classification model of SVM are shown in Table 2.

Table 2. Classification model evaluation results of SVM

	accuracy	recall rate	accuracy	F1
training set	0.725	0.725	0.634	0.668
test set	0.7	0.7	0.553	0.618

Accuracy refers to the proportion of correctly predicted samples to the total samples, recall refers to the proportion of predicted positive samples among the results that are actually positive samples, and precision refers to the proportion of actual positive samples among the results that are predicted to be positive samples. However, because precision rate and recall rate affect each other, although we hope that both the higher the better, but in reality, the two are inversely proportional to each other, the higher the value of one, the other is low, we want to take into account the two, we introduced the F1 indicator, the calculation of precision rate and the recall rate of the reconciliation of the average.

RESULTS ANALYSIS: In this case study, we are more concerned with the prediction of actual positive samples, i.e., the prediction of actual hematoma dilatation. The tolerance of errors such as

no actual hematoma dilatation but predicted as hematoma dilatation is higher, causing less impact; however, the tolerance of errors such as actual hematoma dilatation but predicted as no hematoma dilatation is lower, causing serious impact. Therefore, when selecting the model, we hope that all the indicators are high while paying more attention to the accuracy and recall. By analyzing the prediction evaluation metrics for the training and test sets, we found that the model had the same accuracy and recall for the training and test sets, which were 0.725 and 0.7, respectively, which was the best.

To further ensure that the model is optimal before using it for the prediction of all subsequent patients, we consider using other machine learning classification algorithms to construct classification prediction models, and compare them with the SVM-based classification prediction model, so as to select the optimal model by evaluating the model.

3.6. Constructing classification model prediction probability based on random forests

3.6.1. Introduction to Random Forest Algorithm

Decision tree is a widely used tree prediction model, in the construction process, starting from the root node of the tree, each node selects the optimal attributes for splitting, and keeps constructing the tree until the stopping condition is satisfied [5]. When using the decision tree model, the data needs to be divided into two parts: the training set and the test set. The training set is first trained for modeling to get the classification criteria, and then the test set is used for model evaluation. However, the decision tree also has some shortcomings: (1) the number of nodes increases as the depth of the tree increases, resulting in a decrease in the number of leaf nodes; (2) the construction of the tree will be repeated, resulting in redundancy, which makes the training error larger and causes the model to be overfitted.

Based on the shortcomings of decision trees, the method of random forest is proposed. Random forest is a method based on the decision tree algorithm, using Bagging technology, which is able to generate many classifiers and summarize their results [6-10]. Firstly, a portion of randomly selected samples from the original samples are used to construct a single decision tree. In the node splitting process of each single decision tree, the random selection of feature subspace is used for splitting. Finally, the final prediction results and prediction probability are derived by counting the prediction results of all single decision trees and utilizing the method of voting decision.

3.6.2. Constructing a Random Forest Classification Model

Assuming that the original training set is N and the decision multiple trees built by random forest is M, then M=m_1, m_2, ···m_t resamples the original training set N by bootstrap, i.e., repeat the random sampling of the training set N with put back many times, extract n, and take the extracted data as the new training set.

Constructing a decision tree: each training set is constructed into a decision tree, using sample set 1 as the training sample for constructing decision tree m1, and j to denote any node on m1.

The classification prediction and regression prediction formulas for random forests are as follows:

Categorical Prediction:

$$H(x) = \operatorname{argmax} \sum I\left(h_i(x)\right) = Y \tag{24}$$

Regression prediction:

$$H(x) = \operatorname{avg}\left(\sum h(x_i)\right) \tag{25}$$

The random forest classification prediction model was constructed using spspro, and the model parameters were continuously adjusted to fit the optimal model, which was finally determined to be

modeled with a training set of 0.7 and a test set of 0.3, with put-back sampling and shuffling of the data, and gini coefficients were taken as the evaluation criteria for the node splits. The following outputs are obtained as shown in Figure 5.

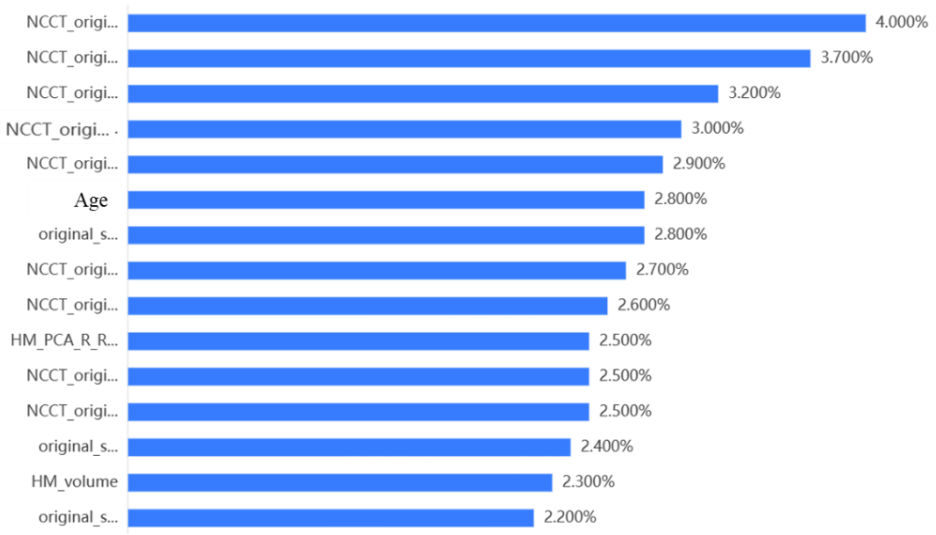

Figure 5. Presentation of partial results of feature importance based on Random Forest classification model

RESULTS ANALYSIS: The three major factors affecting hematoma expansion can be seen from the characteristic importance plot as "NCCT_original_firstorder_Kurtosis", "NCCT_original_firstorder_Range", and "NCCT_original_firstorder_Mean".

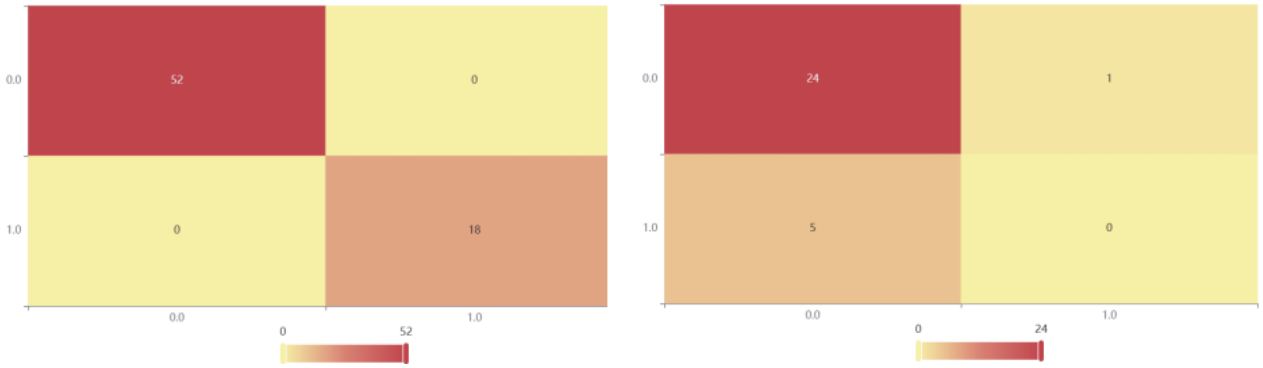

Figure 6. Heat map of confusion matrix for training set (left) and test set (right) of Random Forest classification model

Result analysis: Figure 6 demonstrates the confusion matrix of the test set, in which the number of correct judgments in the training set is 70 and the number of incorrect judgments is 0. The number of correct judgments in the test set is 24 and the number of incorrect judgments is 6. The results of the classification model evaluation for random forest are shown in Table 3.

Table 3. Classification model evaluation results for random forests

	accuracy	recall rate	accuracy	F1
training set	1	1	1	1
test set	0.8	0.8	0.69	0.741

RESULTS ANALYSIS: As mentioned above, in this case study, we were more concerned with the prediction of actual positive samples, i.e., the prediction of actual hematoma dilatation. The tolerance of errors such as no actual hematoma dilatation but predicted hematoma dilatation is high, resulting in a small impact; however, the tolerance of errors such as actual hematoma dilatation but predicted

no hematoma dilatation is low, resulting in a serious impact. Therefore, when selecting the model, we hope that all the indicators are high while paying more attention to the accuracy and recall.

By analyzing the prediction evaluation metrics for the training and test sets, we found that the model has the same accuracy and recall of 0.8 for the test set, with a precision of 0.69 and the best F1 metric of 0.741.

3.7. Model Selection and Prediction

By comparing the classification prediction models constructed by the two machine learning algorithms, we find that the classification prediction model constructed based on Random Forest shows better than the classification prediction model constructed based on SVM in terms of each model evaluation index. At the same time, Random Forest has another advantage that gives the feature importance of each feature indicator, which is conducive to our subsequent analysis, so we chose the classification prediction model constructed based on Random Forest to conduct a prediction analysis on whether hematoma expansion occurs in all patients and calculate the probability value. The predictions of the random forest classification model and their probability values are shown in Table 4.

Table 4. Predictions of the random forest classification model and its probability values are shown as part of the results

Patient Number	Predicted outcome_Y	Probability of predicted outcome_0.0	Predicted Probability of Outcome_1.0
sub001	0	0.9177	0.0823
sub002	0	0.7567	0.2433
sub003	1	0.2314	0.7686
sub004	0	0.9157	0.0843
sub005	0	0.7322	0.2678
sub006	0	0.8638	0.1362
sub007	0	0.9251	0.0749
sub008	0	0.9177	0.0823
sub009	1	0.2237	0.7763
sub010	0	0.8524	0.1476
sub011	0	0.8457	0.1543
sub012	0	0.8158	0.1842
sub013	0	0.7856	0.2144
sub014	0	0.8362	0.1638
sub015	0	0.8985	0.1015

4. Conclusion

In this paper, we conducted a prediction study on the probability of hematoma dilatation in hemorrhagic stroke patients by constructing a classification model based on support vector machine (SVM) and random forest. The results show that the random forest-based classification model has high accuracy in predicting hematoma expansion, which is better than the SVM-based classification model. This finding provides clinicians with a more accurate prediction tool, which helps to identify high-risk patients at an early stage, reduce the incidence of secondary brain injury, and improve the quality of survival and prognosis of patients.

In this study, the data quality was improved in terms of data processing by matching and integrating the patient data and handling the abnormalities, which lays the foundation for the subsequent model construction. In addition, this study also compared the predictive performance of different classification models, providing clinicians with a more targeted basis for decision-making. However,

this study still has some limitations, such as small sample size and single data source, etc. Future studies can further expand the sample size and explore the fusion method of multiple data sources to improve the generalization ability and accuracy of the prediction model.

In conclusion, the random forest-based classification model proposed in this paper has some application value in predicting hematoma expansion in hemorrhagic stroke patients. Through the early identification and intervention of high-risk patients, it is expected to reduce the incidence of secondary brain injury and improve the survival quality and prognosis of patients. Future studies will continue to optimize the model structure with the aim of providing a more accurate prediction tool for clinical diagnosis and treatment.

References

[1] LI Jianbo, GENG Wei, ZHOU Tao, et al. Predictive modeling of hematoma expansion risk in patients with cerebral hemorrhage based on computed tomography angiographic features [J]. Journal of practical clinical medicine, 2022 (026 - 008).

[2] B H Z A, B Z, B Z S, et al. Machine learning-based modified BAT score in predicting hematoma enlargement after spontaneous intracerebral hemorrhage [J]. 2021.

[3] XU Lei, GE Huaizhi, ZHANG Zhijing, et al. The value of predicting hematoma enlargement after spontaneous intracerebral hemorrhage based on clinical features and modeling of imaging histology features in plain CT [J]. Journal of Wenzhou Medical University, 2021, 51 (10): 6.

[4] FAN Jingqian, JIANG Yuyan. Research progress on the mechanism of edema formation around hematoma after cerebral hemorrhage [J]. Shandong Medicine, 2021, 61 (2): 3.

[5] Cortes C, Vapnik V. Support-vector networks [J]. Machine learning, 1995, 20 (3): 273 - 297.

[6] Cui CJ. Research on optical image target recognition technology based on SVM and CNN [D]. Harbin Institute of Technology, 2022.

[7] Ahmed M Y O Z M. Research on privacy-preserving SVM model for distributed medical data analysis [D]. University of Electronic Science and Technology, 2018.

[8] Yin J P. Research on support vector machine kernel function and key parameters selection [D]. Harbin Institute of Technology, 2016.

[9] Li Lin. Disease prediction analysis based on optimized random forest mixture model [D]. North China University of Water Resources and Hydropower, 2022.

[10] Sounak C, Mohammed K, Mihail P. Predicting disease risks from highly imbalanced data using random forest [J]. BMC Medical Informatics and Decision Making, 2011, 11 (1).