Observational Study

Medicine

OPEN

Multimodal ischemic stroke recurrence prediction model based on the capsule neural network and support vector machine

Daying Fan, MM[a], Rui Miao, PhD[b], Hao Huang, PhD[c], Xianlin Wang, MM[a], Siyuan Li, MM[a], Qinghua Huang, MM[b], Shan Yang, MM[a], Renli Deng, PhD[a,*]

Abstract

Ischemic stroke (IS) has a high recurrence rate. Machine learning (ML) models have been developed based on single-modal biochemical tests, and imaging data have been used to predict stroke recurrence. However, the prediction accuracy of these models is not sufficiently high. Therefore, this study aimed to collect biochemical detection and magnetic resonance imaging (MRI) data to establish a dataset and propose a high-performance heterogeneous multimodal IS recurrence prediction model based on deep learning. This is a retrospective cohort study. Data were retrospectively collected from 634 IS patients in Zhuhai, China, a 12-month follow-up was conducted to determine stroke recurrence. We propose the ischemic stroke multi-group learning (ISGL) model, an integrated model for predicting the recurrence risk of multimodal IS in patients, based on a capsule neural network and a linear support vector machine (SVM). Two capsule neural network prediction models based on T1 and T2 signals in the MRI data and a SVM prediction model based on biochemical test data were established. Finally, a vote was conducted on the final judgment of the integrated model. The ISGL model was compared with 6 classical ML and deep learning models: k-nearest neighbors, SVM, logistic regression, random forest, eXtreme Gradient Boosting, and visual geometry group. The results revealed that the accuracy, specificity, sensitivity and the area under the curve of the ISGL model were 95%, 96%, 94%, and 95%, respectively. Among the comparison models, the visual geometry group method exhibited the best performance, but it much lower than those of the ISGL model. Analysis of the importance of biochemical test data revealed that low-density lipoprotein, smoking, and heart disease history were the positively correlated factors, and total cholesterol, high-density lipoprotein, and diabetes were and the negatively correlated factors. This study proposes the ISGL model can be used simultaneously with MRI and biochemical data to predict IS recurrence. This combination resulted in higher rate of performance than that of the other ML models. Additionally, this study found related risk factors affected recurrence, which can be used to intervene in high-risk patients' recurrence as early as possible and promote the development of secondary prevention of stroke.

Abbreviations: AUC = The area under the curve, HDL = high-density lipoprotein, IS = ischemic stroke, ISGL = ischemic stroke multi-group learning, LDL = low-density lipoprotein, ML = machine learning, MRI = magnetic resonance imaging, SVM = support vector machine.

Keywords: deep learning, ischemic stroke, machine learning, predictive model, recurrence, secondary prevention

1. Introduction

Stroke is the second leading cause of death and the third leading cause of disability in the world.[1–3] In China, stroke is the leading cause of death and disability.[4] Patients with recurrent stroke have more severe functional disabilities than those with first-episode stroke.[5] The focus of this study was

DF and RM contributed equally to this article.

The development of an early warning system in cerebral infarction recurrence prediction model based on multi-omics and its implement demonstration. Supported by the Science and Technology Project of Guizhou Province, Project Number: Guizhou Science and Technology Support [2021] General 446, Guangdong Provincial Department of Education Youth Innovative Talent project (No. 2023KQNCX155), Postdoctoral training project of Zunyi Medical University (No. 2023F-ZH-019).

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request.

This study was approved by the Ethics Committee of the Fifth Affiliated Hospital of Zunyi Medical University (Zhuhai), ethical review approval number: [2020] 2020ZH0067. All patients signed a written informed consent form.

[a] Nursing Department, The Affiliated Hospital of Zunyi Medical University, Zunyi, China, [b] Basic Teaching Department, Zhuhai Campus of Zunyi Medical University, Zhu Hai, China, [c] Neurological Department, The Affiliated Hospital of Zunyi Medical University, Zunyi, China.

** Correspondence: Renli Deng, Nursing Department, The Affiliated Hospital of Zunyi Medical University, Zunyi 563000, China (e-mail: dengrenli.research@outlook.com) .*

How to cite this article: Fan D, Miao R, Huang H, Wang X, Li S, Huang Q, Yang S, Deng R. Multimodal ischemic stroke recurrence prediction model based on the capsule neural network and support vector machine. Medicine 2024;103:35(e39217).

Received: 23 January 2024 / Received in final form: 6 March 2024 / Accepted: 17 July 2024

http://dx.doi.org/10.1097/MD.0000000000039217

ischemic stroke (IS), which is the most common type of stroke, accounting for approximately 80% of stroke cases.[6] Studies have revealed that the secondary prevention of stroke can reduce the risk of IS recurrence events by approximately 13% and up to 67%.[7] The accurate identification of risk factors is the premise and foundation of secondary prevention and is the most effective means of reducing the disability and mortality rates of patients with IS.[8] Therefore, it is important to establish a high-performance and comprehensive IS recurrence model and identify possible factors influencing stroke recurrence in patients.

In clinical practice, senior neurologists combine patient baseline data, laboratory tests, and imaging examinations to comprehensively develop personalized diagnoses and treatment and nursing programs.[9,10] However, no study has constructed a multimodal IS recurrence dataset to develop a prediction model. Most existing studies were based on traditional statistical methods for analyzing single-modal biochemical test data, which exhibit simple calculation methods and poor prediction effect, leading to less comprehensive and robust clinical applications. Recently, artificial intelligence has flourished in the field of cerebrovascular disease.[11–13] Machine learning (ML) methods have better sensitivity and specificity for screening test data features and identifying image features.[14,15] Many studies have shown that compared with traditional statistical methods, ML methods can be effectively applied to IS recurrence prediction and can better predict results.[16,17] Therefore, this study aimed to use multimodal data to construct a deep learning model for IS recurrence.

Baseline, biochemical test, and magnetic resonance imaging (MRI) data were collected for the following purposes. First, a multimodal IS recurrence feature dataset was established based on the patients' complete clinical data. Second, a heterogeneous multimodal IS recurrence risk prediction ensemble model was proposed, based on the capsule neural network and linear support vector machine (SVM) methods (Fig. 1). This model integrates multisource heterogeneous data of patients with IS and analyzes the importance of the biochemical test data. Finally, the 6 most popular ML algorithms were compared with the ischemic stroke multi-group learning (ISGL) model to verify their prediction performance. In addition, the feature importance of the model was analyzed, and 6 important biochemical detection features that were most related to stroke recurrence were identified.

2. Material and methods

2.1. Study design and study population

This was a single-center, retrospective cohort study, and all data were obtained from the hospital's electronic health record system. All subjects were recruited at the Fifth Affiliated Hospital of Zunyi Medical University (Zhuhai) from June 1, 2017, to June 30, 2019. In total, 634 patients were recruited (Fig. 2). The inclusion criteria were: (1) patients aged ≥18 years; (2) patients diagnosed with IS at discharge (i.e., diagnostic criteria of various cerebrovascular diseases in the 2015 edition of the Chinese Classification of Cerebrovascular Diseases); and (3) patients who had MRI on admission or during hospitalization (no surgical or drug treatment was performed). The exclusion criteria were as follows: (1) patients who had died or were transferred to the hospital; (2) patients with a

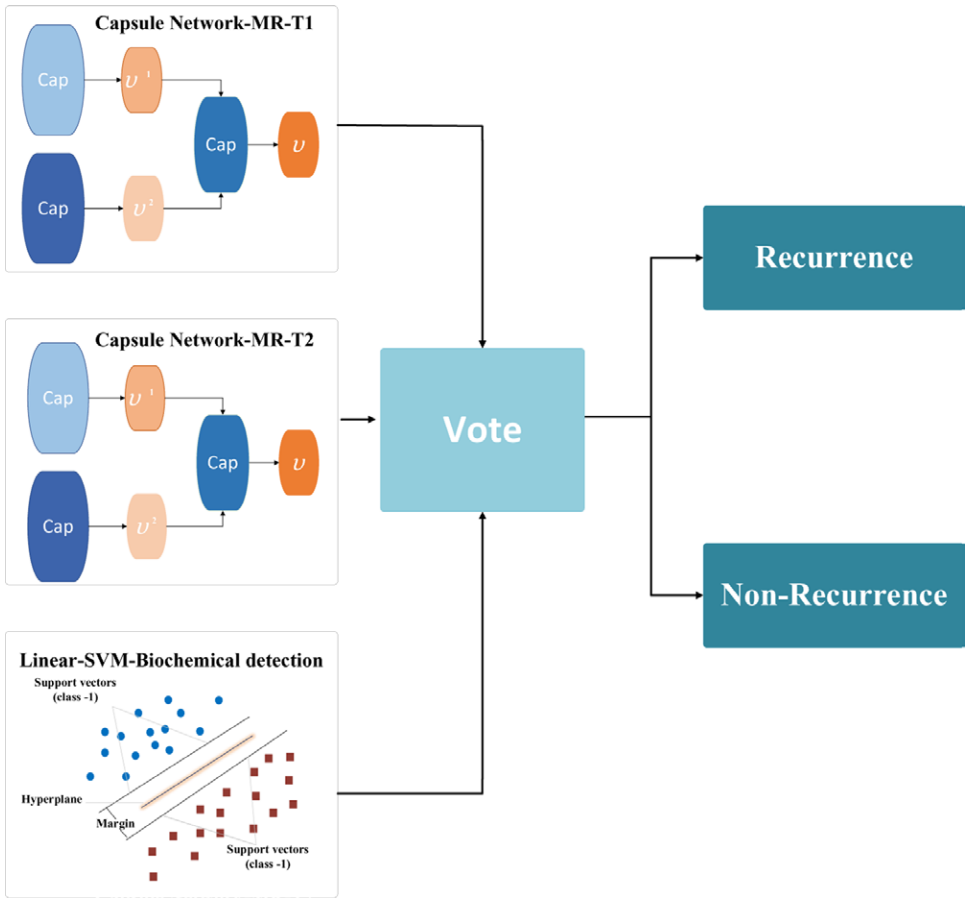

Figure 1. ISGL model constructed by the capsule neural network (i.e., deep learning model), the linear-SVM model (i.e., ML model), and the voting method used to judge the final recurrence result. ML = machine learning.

history of stroke; (3) patients diagnosed with transient ischemic attack, cerebral embolism, cerebral watershed infarction, or other definite etiologies of IS; and (4) patients in whom the endpoint event (i.e., recurrence) could not be determined. All procedures involving human participants were consistent with the Declaration of Helsinki (revised in 2013 by the reference[18]).

2.2. Feature selection

To determine the biochemical test collection indicators related to the risk of IS recurrence and to use these variables to create a prediction model, a literature review and word frequency analysis were conducted to obtain the recurrence factors of the modal test data. First, the risk factors for IS recurrence were extracted by searching the literature related to risk factors for IS recurrence over the past 10 years (Fig. 3). Subsequently, the Jieba Library in Python software was used to complete the word frequency analysis of the risk factors, and 78 recurrence risk factors were extracted. Based on the results of the frequency analysis and availability of hospital-related examinations, 30 recurrence risk factors were included as the collection content of the modal test data. Demographic data (e.g., sex, age, and length of hospital stay), medical history (e.g., smoking, drinking, history of peptic ulcers, hypertension, diabetes, heart disease, and atherosclerosis), self-care scores, and hospitalization event information (e.g., recurrence or not) were collected from the patients' electronic medical records. This study also recorded the laboratory examination data of the patients on admission, including uric acid, triglyceride, total cholesterol, high-density lipoprotein (HDL), low-density lipoprotein (LDL), homocysteine, C-reactive protein, serum albumin level, apolipoprotein A1, apolipoprotein B, platelet level, white blood cell level, neutrophil percentage, fibrinogen, and glycosylated hemoglobin. Two basic MRI images were collected (the patients did not undergo surgery nor were they treated with medication during the MRI examination): T1-weighted imaging and

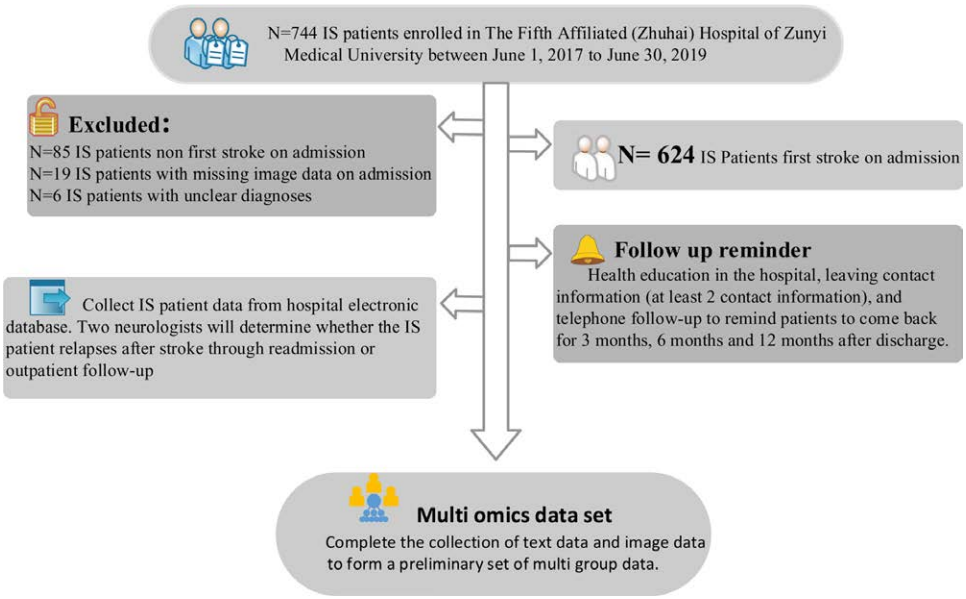

Figure 2. Flow chart of patient recruitment.

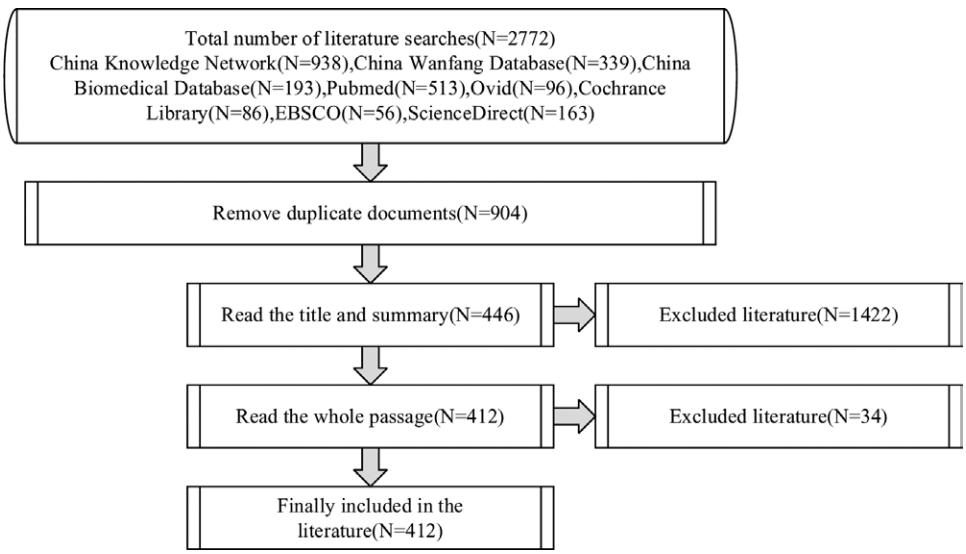

Figure 3. Flow chart of literature screening.

T2-weighted imaging, as the model features of the imaging modality.

2.3. Outcome definitions

Before the patients were discharged from the hospital, they were given corresponding health education and asked to provide the researchers their phone numbers and those of at least 2 family members. The patients were reminded by telephone to return for follow-up visits at 3, 6, and 12 months post discharge. Two neurologists assisted with these procedures. According to a report by the World Health Organization, comprehensive clinicopathological information, and computed tomography and/or MRI are required to diagnose IS during hospitalization.[19] The diagnostic criteria were as follows: (1) new neurological deficit symptoms appearing after the symptoms and signs of the original neurological deficit improved or disappeared; (2) new ischemic lesions confirmed by a head computed tomography and/or MRI; and (3) exclusion of progressive stroke and disease progression.

2.4. Sample size

According to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines, the rule of thumb for sample size is that each variable has at least 10 outcome events.[20] According to the risk factors for recurrence included in the modal test data, 31 variables were incorporated into the construction of the ISGL model. This study required at least 310 patients with first-episode IS to be recruited for the model development.

2.5. Processing of missing data

In the dataset, the original test data contained a small number (2.07%) of vacancy values. The gradient boosting decision tree method was used to predict the vacancy values (i.e., missing-value interpolation) based on known test information. The gradient boosting decision tree, also known as a multiple additive regression tree, is an iterative decision tree algorithm, which constructs a set of weak learners (i.e., trees) and accumulates the results of multiple decision trees as the final prediction output. The algorithm effectively combines the decision tree with the integration of ideas to scientifically and effectively fill in missing data.

2.6. Heterogeneous multimodal model development

In total, 634 patients with first-episode IS were included in this study. Owing to the data imbalance, complete IS recurrence patient data were used for each independent random experiment and 2-fold data of non-recurrence patient data were randomly selected. The test and training sets were then randomly cut using a ratio of 8:2. We performed 100 random repeat experiments on the model and calculated the average value to reduce data overfitting.

2.7. MRI image processing model based on the capsule neural network

The recurrence of IS is closely related to the infarction location; therefore, it is necessary to track the lesion area and stroke infarction signal using imaging examinations. As the occurrence, development, and recurrence of IS is a complex pathological process, the characteristics of the entire image are more representative, and contribute to the prediction of recurrence in patients. Therefore, in this study, we used 3 dimensions images of the entire brain region of a patient for modeling and performed an overall analysis of all plane slice images of the patient's head. The entire spatial structure was incorporated into the model, including the infarct area and its location. Capsule neural networks, which use scalars to record local image information, have been proposed to solve the problems of convolutional neural networks. However, other local information cannot be analyzed with a convolutional neural networks, which renders it difficult to explain the relationship between certain features and the overall image. The capsule neural network uses vector feature states to represent feature information. Vector features no longer represent the specific existence of local features, but different properties of the same global image.[21] Currently, there is no research on the construction of stroke recurrence prediction models using capsule networks. In previous studies, the capsule network had a positive effect on MRI image processing. Researchers have used a capsule neural network to segment MRI images, which effectively improved the cutting accuracy.[22] More researchers have realized automatic segmentation of the left ventricle in cardiac MRI based on a deep learning model of the capsule network.[23] The MRI data in this study were modeled using a capsule neural network, which accurately captured the relationship between infarction location and IS recurrence.

The capsule neural network uses vector feature states to represent feature information. Vector features no longer represent the specific existence of local features but different properties of the same global image. Capsule network vector has the following properties: 1. The modulus of the vector represents the probability of the existence of the feature; 2. The direction of the vector represents the attitude information of the feature; 3. Moving features will change the Capsule vector, which does not affect the probability of feature existence. The general overall operation mode of the capsule network is as follows:

Enter 2 vectors, v^1 and v^2, multiplied by the weight matrix w^1 and w^2, get and u^1 and u^2:

$$u^1 = w^1 v^1$$

$$u^2 = w^2 v^2$$

Next, calculate s by the following formula,

$$s = c_1 u^1 + c_2 u^2$$

Subsequently, the squash activation function is used to obtain v;

$$v = \text{Squash}\,(s)$$

$$v = \frac{\|s\|^2}{1 + \|s\|^2} \quad \frac{s}{\|s\|}$$

Among them, the calculation process of c_1 and c_2 becomes a dynamic routing process, and its workflow is as follows:

$$b_1^0 = 0, b_2^0 = 0$$

$$\text{For} \quad r = 1 \quad \text{to} \quad T \quad \text{do}$$

$$s^r = c_1 u^1 + c_2 u^2$$

$$a^r = \text{Squash}\,(s^r)$$

$$b_i^r = b_i^{r-1} + a^r * u^i$$

2.8. Biochemical detection data processing model based on the support vector machine

An SVM is a generalized linear classifier that classifies data using supervised learning. The decision boundary is the maximum

Table 1

Characteristics of the IS patients. For continuous data, values are expressed as medians in IQR (quartile range). Other values are expressed in numbers and percentages.

	Category/company	Total (N = 634)	Recurrence (N = 77)	No recurrence (N = 557)	*P* value
Gender (N/%)	Male	364 (57.4)	49 (63.6)	315 (56.6)	.269
Age (IQR)	Year	69 (59,77)	72 (65.5,78)	68 (59,77)	.038
Length of hospitalization (IQR)	Day	7 (5.75,9)	8 (6,9.5)	7 (5,9)	.026
Smoking history (N/%)	Yes	137 (21.6)	30 (39.0)	107 (19.2)	<.000
Drinking history (N/%)	Yes	106 (16.7)	23 (29.9)	83 (14.9)	.002
History of peptic ulcers (N/%)	Yes	68 (10.7)	8 (10.4)	60 (10.8)	1.000
Diabetes (N/%)	Yes	136 (21.4)	20 (26.0)	116 (20.8)	.302
Hypertension (N/%)	No	190 (30.0)	13 (16.9)	177 (31.8)	.008
	Primary	41 (6.4)	4 (5.2)	37 (6.6)	
	Secondary	72 (11.4)	11 (14.3)	61 (11.0)	
	Tertiary	331 (52.2)	49 (63.6)	282 (50.6)	
Admission microcomputer blood glucose (IQR)	mmol/L	7 (6.1,8.7)	7.3 (6.1,9.3)	6.9 (6,8.6)	.053
Heart disease (N/%)	Yes	94 (14.8)	16 (20.8)	78 (14.0)	.124
Atrial fibrillation (N/%)	Yes	29 (4.6)	7 (9.1)	22 (3.9)(96.1)	.072
Admission systolic blood pressure (IQR)	mm Hg	150 (136,169)	153 (141,176.5)	149 (135,167)	.047
Admission diastolic blood pressure (IQR)	mm Hg	86 (77,94)	87 (79.5,97)	86 (76,94)	.207
Self-care (N/%)	Fully self-care	237 (37.4)	25 (32.5)	212 (38.1)	<.000
	Need care	397 (62.6)	52 (67.5)	345 (61.9)	
Carotid atherosclerosis (N/%)	Yes	550 (86.8)	75 (97.4)	475 (85.3)	<.000
Uric acid (IQR)	μmol/L	344.5 (284,402.3)	359 (298.5407.5)	343 (279,399.5)	.140
Triglyceride (IQR)	mmol/L	1.34 (0.99,1.91)	1.44 (1.07,1.89)	1.33 (0.99,1.92)	.319
Total cholesterol (IQR)	mmol/L	4.72 (4.11,5.27)	4.77 (4.13,5.2)	4.71 (4.1,5.28)	.715
High-density lipoprotein (IQR)	mmol/L	1.16 (0.97,1.39)	1.15 (0.98,1.37)	1.16 (0.97,1.39)	.709
Low-density lipoprotein (IQR)	mmol/L	3.14 (2.54,3.65)	3.21 (2.81,3.68)	3.13 (2.53,3.66)	.301
Homocysteine (IQR)	umol/L	9.1 (7.6,11.2)	9.7 (8.4,11.6)	8.9 (7.5,11.2)	.014
C-reactive protein (IQR)	mg/L	2.45 (0.93,7.71)	2.57 (0.97,8.77)	2.42 (0.92,7.59)	.877
Serum albumin (IQR)	g/L	40.4 (38.2,42.7)	40 (37.6,41.8)	40.5 (38.3,42.8)	.085
Apolipoprotein A1 (IQR)	g/L	1.08 (0.98,1.17)	1.09 (0.99,1.15)	1.08 (0.98,1.18)	.729
Apolipoprotein B (IQR)	g/L	1.02 (0.87,1.15)	1.04 (0.92,1.15)	1.02 (0.87,1.15)	.398
Platelet count (IQR)	L	227 (194,256)	227 (195,256)	227 (194,256)	.940
Glycosylated hemoglobin (IQR)	%	5.9 (5.6,6.5)	6.1 (5.8,7)	5.9 (5.7,6.4)	.823
Fibrinogen (IQR)	L	3.17 (2.75,3.59)	3.23 (2.83,3.73)	3.16 (2.74,3.57)	.420
Leukocyte count (IQR)	L	7 (5.7,8.3)	7.1 (6.2,8)	7 (5.7,8.3)	.140
Neutrophil percentage (IQR)	%	64.2 (57.8,70.2)	63.3 (56.7,69.2)	64.3 (57.9,70.3)	.004

margin hyperplane for the learning samples. This method has been widely used in various fields. The core idea is to solve a hyperplane formula for a two-dimensional classification problem. In the SVM model, a linear function was chosen as the kernel function. That is to solve the following formula:

$$W^T X + b = 0$$

2.9. Model voting

Based on the imaging and test data collected, 3 sub-models of multimodal IS recurrence prediction were constructed based on the capsule neural network (i.e., deep learning model) and the linear-SVM model (i.e., ML model). Voting is an ensemble learning model that follows the principle of majority. The overall classification performance of the model was improved through the integration of multiple models and the error rate of the model was reduced. Ideally, the prediction effect of the voting method would be better than that of any sub-model. The voting method was chosen to integrate the prediction results of the 3 sub-models and construct the ISGL model.

2.10. Model performance evaluation

The area under the curve (AUC), receiver operating characteristic curve, accuracy, specificity, and sensitivity were used to evaluate the predictive performance of the ISGL model. Six algorithms—k-nearest neighbors, SVM, random forest, logistic regression, eXtreme Gradient Boosting, and visual geometry group—were used as benchmark models for comparison with the ISGL-integrated model.

2.11. Statistical analysis

The data were preprocessed before constructing the modal test data model. The continuous variables in these data did not conform to a normal distribution; therefore, they were expressed as median and quartile ranges and analyzed using the Mann–Whitney *U* test. Categorical variables were expressed as percentages and analyzed using the Pearson chi-square test. All statistical tests were two-tailed, and statistical significance was set at $P < .05$. The algorithms used in the study were extracted from the Python 3.7 and SPSS (version 29.0) software.

3. Results

3.1. Baseline characteristics

Overall, 634 adult patients met the inclusion criteria, and 77 patients relapsed within 1 year after their first-episode stroke, with a recurrence rate of 12.1%. The multimodal IS recurrence prediction dataset established in this study primarily contained 3 types of modal data: modal test, MRI T1 signal, and MRI T2 signal data. The data used in this study contained 31 features. Of the 634 patients, 57.4% were male. The median patient age was 69 years (interquartile range = 59–77). The 3 most common comorbidities were carotid atherosclerosis (86.8%), hypertension (70%), and diabetes (21.4%). Table 1

shows the feature distributions of the patient-relevant test variables.

3.2. Performance of the ISGL model with 6 popular algorithms

The accuracy, specificity, and sensitivity of the ISGL model are 95%, 96%, and 94%, respectively. The test results were in line with the expectations. The results of the integrated model were superior to those of the single-modal image data and modal test data, and the AUC reached 95% (95% CI, 0.94–0.96). The image data revealed very good test results (92% and 93% accuracy), whereas the text modal data results were relatively poor (69% accuracy), in line with previous experimental results. The integrated model results were superior to the single-image modal and text modal data results (95% accuracy). This indicated that the integrated ISGL model was superior to the single-modal IS prediction model (Table 2). The decision, calibration, and PR-AUC curves of the ISGL model are shown in Figures 4, 5, and 6. The decision curve reveals that the net income value of the model is satisfactory, and the calibration curve reveals that the model's prediction and observation probabilities are highly consistent. The PR-AUC curve revealed that the ISGL model had a superior ability to correctly predict true lesions. The performances of 6 popular algorithms on the test set are shown in Figure 7. The ISGL model performed well in the experimental results, and the F1-score was much larger than that of the other 6 algorithms, which indicates that the overfitting of the model was alleviated. Among the 6 algorithms, the model with the best prediction performance was visual geometry group, with an accuracy, specificity, sensitivity, AUC, and F1-score of 0.76, 0.79, 0.47, 0.87, and 0.47, respectively (Table 3).

3.3. Feature weight of the modal test data

Finally, 30 modal test data variables were included in the construction of the ISGL model, including the baseline data, clinical history, and biochemical test data of patients with IS. The weight values of the 30 variables were calculated by constructing a linear SVM model for the modal test data. Finally, 18 positive and 12 negative risk factors affecting recurrence of first-episode IS were identified (Figs. 8 and 9). Among them, for the positive risk factors, the weight values of LDL (1.25343591), smoking (1.17272469), and history of heart disease (1.05821598) were >1, whereas for the negative risk factors, the weight value of total cholesterol (1.1022789) was >1, followed by HDL (0.19130034) and diabetes (0.12935474).

4. Discussion

Eligible patients were recruited to construct an IS recurrence model. By collecting complete baseline, laboratory, and imaging examination data from the hospital information system, a

Table 2

Comparison of the accuracy of the integrated ISGL model and the independent modeling of the 3 modes.

	Accuracy (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	AUC (95% CI)
Textomics	69% (68–70%)	70% (68–71.5%)	67% (66–68%)	69% (68–70%)
MRI T1	93% (92.5–94%)	94% (93–95%)	92% (90.5–93.5%)	93% (92–94%)
MRI T2	92% (91–93%)	93% (91.5–94.5%)	91% (90–92%)	92% (91–93%)
ISGL model	95% (94–96%)	96% (95–97%)	94% (93–95%)	95% (95–97%)

AUC = The area under the curve.

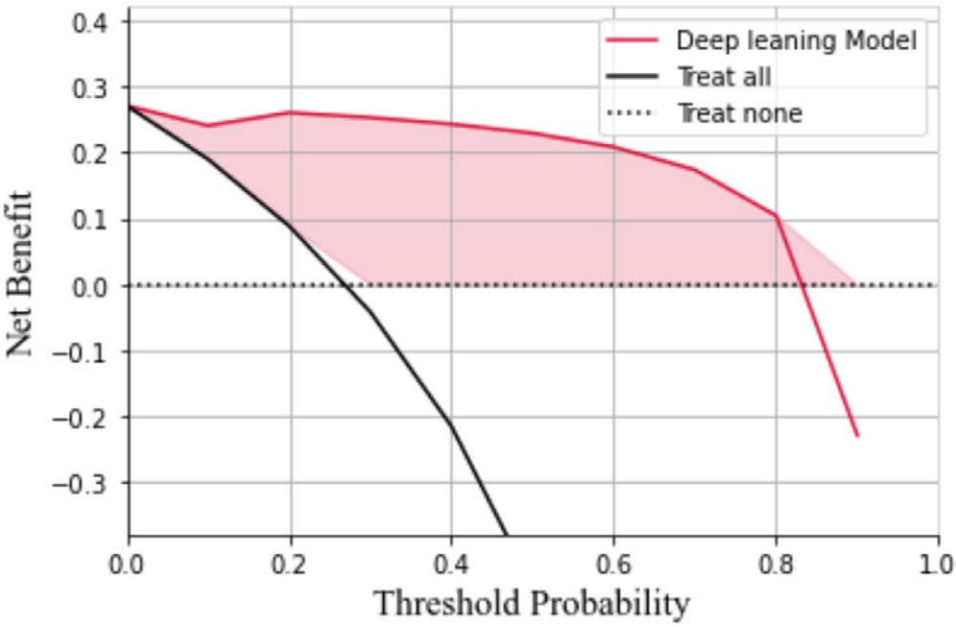

Figure 4. Decision curve of the ISGL model. Dotted line: net benefit of predicting no patients; black line: net benefit of predicting all patients; red line: net benefit of predicting patients according to the ISGL model. The ISGL model-based decisions were supported in the range of threshold probabilities of approximately 0% to 85%. ISGL = ischemic stroke multi-group learning.

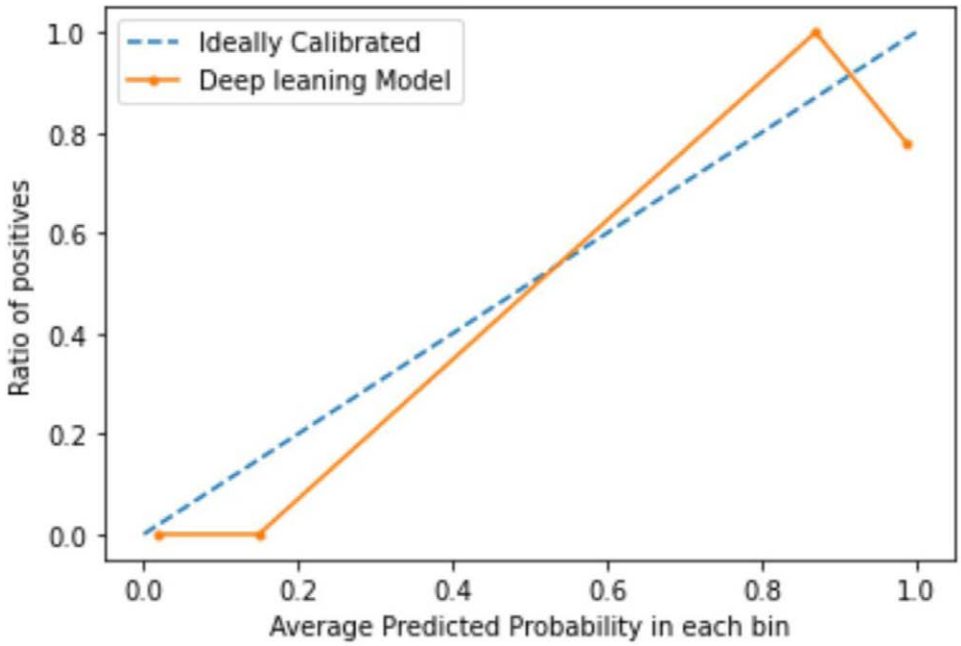

Figure 5. Calibration curve of the ISGL model. The solid line represents the performance of ISGL model; a closer fit to the diagonal dotted line represents a better prediction. ISGL = ischemic stroke multi-group learning.

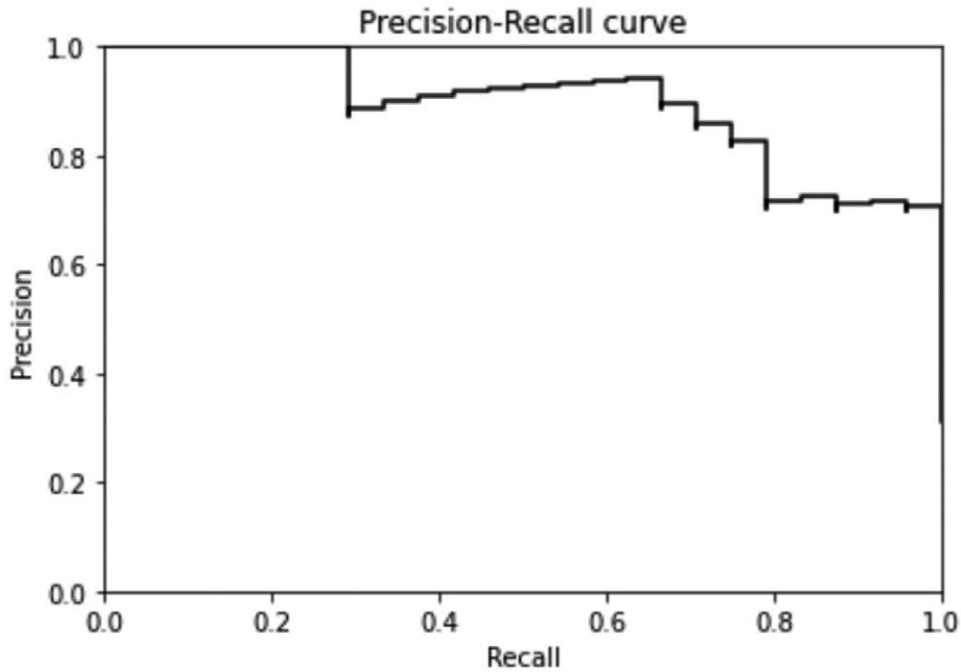

Figure 6. The PR-AUC curve of the ISGL model. The greater the area under the curve (AUC), the better the model was at correctly predicting true lesions. Precision is equivalent to positive predictive value and recall is equivalent to sensitivity. ISGL = ischemic stroke multi-group learning.

multimodal dataset, including test and image modes, was established. The ISGL integrated model was constructed using an SVM, capsule neural network, and model voting method. Six typical ML algorithms were used for comparison with the ISGL model. Regarding the performance of the single-modal data of the ISGL model, the accuracy of MRI-T1 was the highest, followed by MRI-T2, and finally, the biochemical detection data of text model. MRI has a different emphasis on the pathological observation of stroke.[24] T1-weighted imaging can better show anatomical structures, whereas T2-weighted imaging can better show tissue lesions. Based on the current results, MRI-T1 has a more evident predictive effect on IS recurrence. Although the accuracy of text data, such as biochemical tests, is not high, it contains a series of other data, such as the patients' medical history and laboratory tests. This has a complementary and auxiliary effect on the image modality. As a result, the data of the single modality are fused using the voting method, which significantly improves the overall prediction result value.

Previous studies have used traditional statistics to predict IS recurrence. For example, Yu et al (2021) combined logistic regression and laboratory test data to develop a nomogram for predicting IS recurrence in hospitals. The AUC of the nomogram in the validation cohort was 0.717.[25] Compared with test data, such as ordinary laboratory tests, imaging is more valuable for

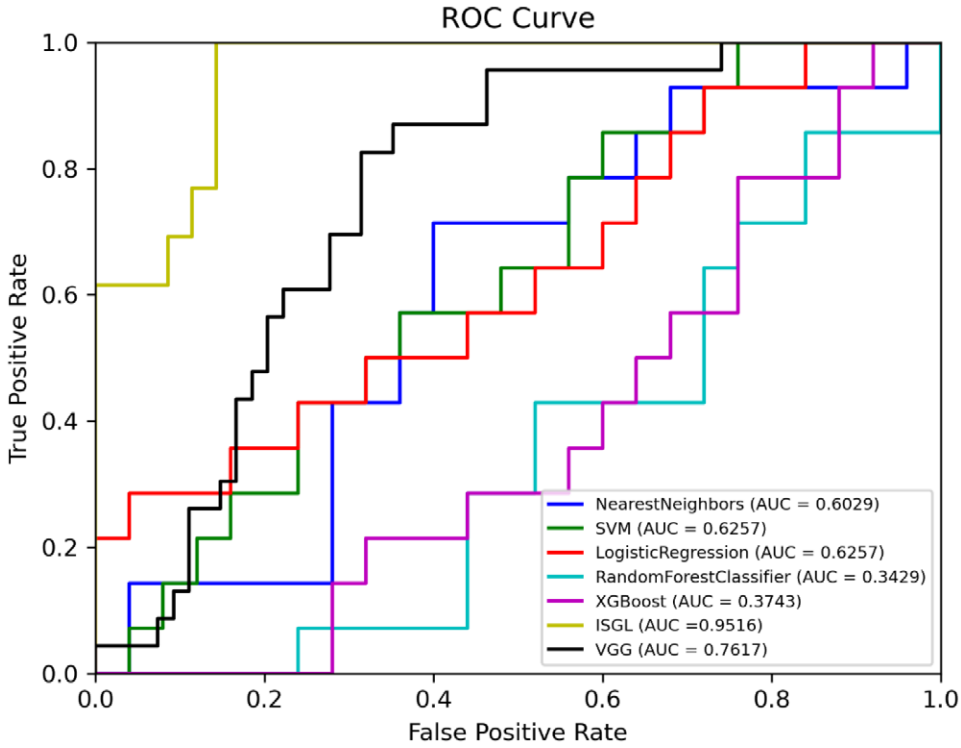

Figure 7. Comparison of the ROC curves between the ISGL and the 6 algorithms. The larger the area under the ROC curve is, the better the prediction performance of the model is. The area under the ROC curve of the ISGL model is larger than that of the 6 machine learning algorithms. ISGL = ischemic stroke multi-group learning, ROC = receiver operating characteristic curve.

Table 3

Accuracy, specificity, sensitivity, AUC, and F1-score values of the ISGL model and 6 popular algorithms.

Algorithm	AUC (95% CI)	Accuracy (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	F1-score (95%CI)
ISGL model	0.9516 (0.95–0.97)	0.95 (0.94–0.96)	0.96 (0.95–0.97)	0.94 (0.93–0.95)	0.92 (0.91–0.93)
KNN	0.6029 (0.58–0.61)	0.56 (0.55–0.57)	0.60 (0.59–0.61)	0.72 (0.70–0.74)	0.62 (0.61–0.63)
SVM	0.6257 (0.59–0.65)	0.63 (0.61–0.65)	0.67 (0.66–0.08)	0.66 (0.65–0.68)	0.72 (0.71–0.73)
Logistic regression	0.6257 (0.59–0.65)	0.64 (0.62–0.66)	0.64 (0.62–0.66)	0.62 (0.61–0.63)	0.45 (0.44–0.46)
Random forest	0.3429 (0.33–0.35)	0.69 (0.67–0.71)	0.20 (0.19–0.21)	0.86 (0.85–0.87)	0.25 (0.24–0.26)
XGBoost	0.4743 (0.45–0.50)	0.62 (0.61–0.63)	0.40 (0.39–0.41)	0.69 (0.67–0.71)	0.35 (0.34–0.36)
VGG	0.7617 (0.74–0.77)	0.79 (0.77–0.81)	0.77 (0.75–0.79)	0.78 (0.76–0.81)	0.77 (0.76–0.78)

KNN = k-nearest neighbors, VGG = visual geometry group, XGBoost = eXtreme Gradient Boosting.

predicting recurrence.[26–28] Therefore, a comprehensive multidimensional prognostic prediction model for stroke, based on imaging and biological markers, has become a new research direction in recent years.[29] Zhang et al used logistic regression to synthesize clinical and imaging information, with an AUC value of 0.753.[1] Owing to the lack of robustness of the logistic regression, single-modal and multimodal data did not show good prediction performance. In recent years, owing to the excellent performance of artificial intelligence in the field of image processing, researchers have combined ML with modal image data to develop models. For example, Lee et al (2020) developed 3 models using 3 ML methods to predict stroke development time, which performed significantly better than manual predictions.[30] In addition, ML achieved excellent performance in modal test prediction. Some studies have used prehospital data to develop ML models to predict large-vessel occlusion in IS, with an AUC value of 0.831.[31] It is generally believed that comprehensive clinical features contribute to the accurate prediction of IS recurrence. Complete baseline data, laboratory test results, and MRI scans of the patients with IS were collected. Therefore, the multimodal ISGL model based on ML has the highest application value for predictions.

In China, patients with IS receive treatment and nursing management from community hospitals at home following discharge; however, community doctors and nurses cannot make the same professional judgments as those made by neurologists. This is detrimental to the prevention and management of recurrence, greatly delaying the treatment of patients and resulting in a higher prevalence of recurrence.[32] In the post-discharge management of stroke, various biochemical tests play important roles in the prevention of stroke recurrence. This study found that LDL level, smoking, and a history of heart disease were the most important positive correlation factors affecting IS recurrence. By managing the indicators (e.g., blood lipids, blood pressure, and blood glucose) of patients, the probability of recurrence and death after a first-episode IS can be effectively reduced. This finding is consistent with that of previous studies, which reported that elevated LDL levels, long-term smoking, and a history of heart disease increased the risk of IS recurrence.[33–35] More attention should be paid to the top 3 negative risk factors, total cholesterol, HDL, and diabetes, which are the most important factors for IS recurrence. HDL, LDL, and total cholesterol are the basic parameters of blood lipid examination; therefore, the

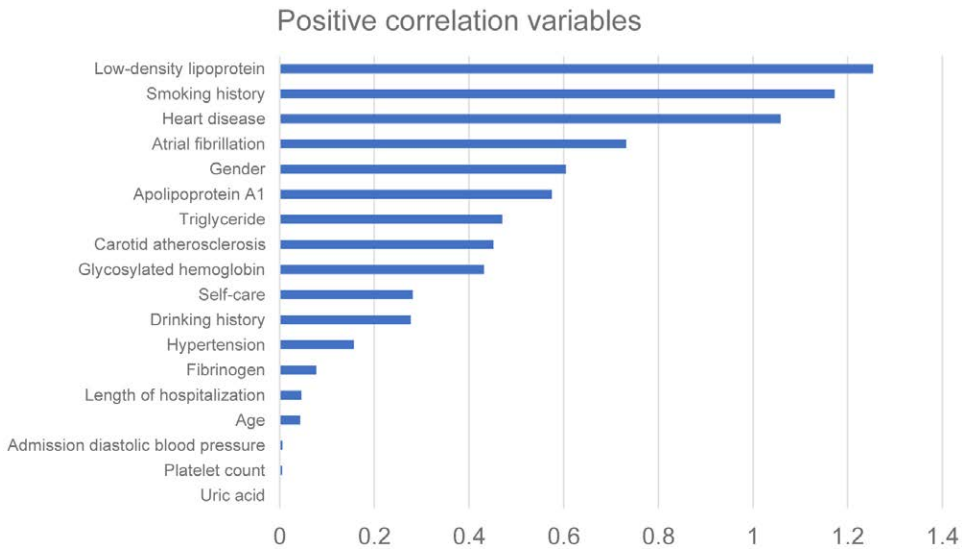

Figure 8. Positive risk factors affecting the recurrence of first-episode IS. IS = ischemic stroke.

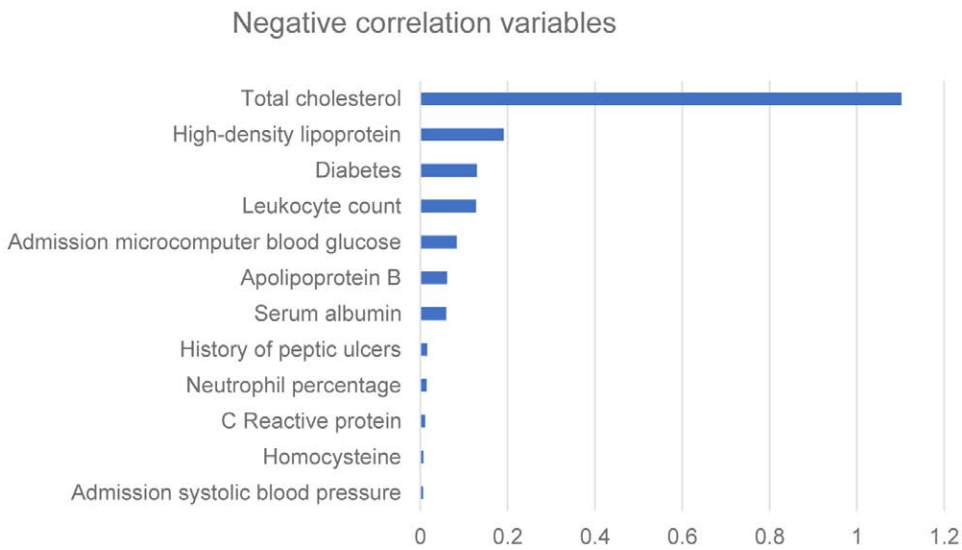

Figure 9. Negative risk factors affecting the recurrence of first-episode IS. IS = ischemic stroke.

control of blood lipids is key to preventing IS recurrence.[36,37] A history of diabetes was negatively correlated with IS recurrence in this study. In a recent study, Wang et al found that stress hyperglycemia is a biomarker of stroke recurrence independent of previous diabetes.[38] In contrast, researchers have found that diabetes is an important risk factor for IS recurrence in women.[39] Considering the different conclusions of various studies, more studies are needed to explore sex differences and the pathological relationship between diabetes and IS recurrence. In general, the information weight of each test variable provided by the ISGL model can help community doctors to implement targeted preventive measures according to the importance of risk factors, as well as help them to better manage stroke secondary prevention and reduce stroke recurrence.[40] The ISGL model constructed in this study exhibited a considerably higher accuracy than that of previous similar IS recurrence prediction models.[41] The ISGL model identified the comprehensive risk of recurrence in patients and provided evidence for its high performance in the secondary prevention of stroke.

This study has few limitations. First, our research was a single-center study, and external verification and promotion of the ISGL prediction model require multicenter testing. Second, in our text modal data collection indicators, there was a lack of national institutes of health stroke scale, Modified Rankin Scale scores, etc, which may have had an impact on the prediction results of the model. Finally, to synthesize the predictive effect of each mode on the model, we did not weigh the data for the different modes. In future work, we plan to expand and improve our database in Zunyi City, Guizhou Province, China, and improve the content of the text data collection indicators in the database. In addition, a multi-center model verification plan was implemented to ensure diversification of the data. Using an more completely diversified dataset, we plan to further improve the model according to the results of different omics models to dynamically weigh different omics and explore the predictive value of different modalities for IS recurrence.

5. Conclusion

In this study, the ISGL model was proposed to predict IS recurrence using both MRI and biochemical test data. The results showed higher prediction performance in comparison with other ML models. In addition, the study also found related risk

factors that affected recurrence, which should be used to intervene in patients with high risk as early as possible and promote the development of the secondary prevention of stroke.

Author contributions

Conceptualization: Rui Miao, Siyuan Li, Qinghua Huang, Shan Yang, Renli Deng.
Data curation: Xianlin Wang, Siyuan Li, Qinghua Huang, Shan Yang.
Formal analysis: Daying Fan.
Funding acquisition: Renli Deng.
Investigation: Rui Miao.
Methodology: Daying Fan, Rui Miao.
Resources: Daying Fan, Rui Miao, Hao Huang, Xianlin Wang, Qinghua Huang.
Software: Rui Miao, Hao Huang, Xianlin Wang, Siyuan Li.
Supervision: Hao Huang, Siyuan Li.
Validation: Rui Miao, Hao Huang.
Visualization: Hao Huang.
Writing – original draft: Daying Fan.
Writing – review & editing: Renli Deng.

References

[1] Zhang K, Fang Y, Fan H, et al. A nomogram for predicting the in-hospital risk of recurrence among patients with minor non-cardiac stroke. Curr Med Res Opin. 2022;38:487–99.
[2] Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics—2014 update: a report from the American Heart Association. Circulation. 2014;129:e28–e292.
[3] Collaborators GBDN. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019;18:459–80.
[4] GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1736–88.
[5] Xing L, Lin M, Du Z, et al. Epidemiology of atrial fibrillation in northeast China: a cross-sectional study, 2017-2019. Heart. 2020;106:590–5.
[6] Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. Circulation. 2015;131:e29–322.
[7] Ikeme JC, Pergola PE, Scherzer R, et al. Cerebral white matter hyperintensities, kidney function decline, and recurrent stroke after intensive blood pressure lowering: results from the secondary prevention of small subcortical strokes (SPS 3) trial. J Am Heart Assoc. 2019;8:e010091.
[8] Algra A, Wermer MJ. Stroke in 2016: stroke is treatable, but prevention is the key. Nat Rev Neurol. 2017;13:78–9.
[9] Subudhi A, Dash P, Mohapatra M, Tan RS, Acharya UR, Sabut S. Application of machine learning techniques for characterization of ischemic stroke with MRI images: a review. Diagnostics (Basel). 2022;12:2535.
[10] Ho KC, Speier W, Zhang H, Scalzo F, El-Saden S, Arnold CW. A machine learning approach for classifying ischemic stroke onset time from imaging. IEEE Trans Med Imaging. 2019;38:1666–76.
[11] Sung SM, Kang YJ, Cho HJ, et al. Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms. Clin Neurol Neurosurg. 2020;195:105892.
[12] Guan W, Ko D, Khurshid S, et al. Automated electronic phenotyping of cardioembolic stroke. Stroke. 2021;52:181–9.
[13] Ong CJ, Orfanoudaki A, Zhang R, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One. 2020;15:e0234908.
[14] Sheth SA, Lopez-Rivera V, Barman A, et al. Machine learning-enabled automated determination of acute ischemic core from computed tomography angiography. Stroke. 2019;50:3093–100.
[15] Zhu H, Jiang L, Zhang H, Luo L, Chen Y, Chen Y. An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging. Neuroimage Clin. 2021;31:102744.
[16] Abedi V, Avula V, Chaudhary D, et al. Prediction of long-term stroke recurrence using machine learning models. J Clin Med. 2021;10:1286.
[17] Fernandez-Cadenas I, Mendioroz M, Giralt D, et al. GRECOS project (Genotyping Recurrence Risk of Stroke): the use of genetics to predict the vascular recurrence after stroke. Stroke. 2017;48:1147–53.
[18] World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310:2191–4.
[19] Hatano S. Experience from a multicentre stroke register: a preliminary report. Bull World Health Organ. 1976;54:541–53.
[20] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Bmj. 2015;350:g7594.
[21] Mazzia V, Salvetti F, Chiaberge M. Efficient-CapsNet: capsule network with self-attention routing. Sci Rep. 2021;11:14634.
[22] Cao YJ, Wu S, Liu C, et al. Seg-CapNet: a capsule-based neural network for the segmentation of left ventricle from cardiac magnetic resonance imaging. J Comput Sci Technol. 2021;36:323–33.
[23] He Y, Qin W, Wu Y, et al. Automatic left ventricle segmentation from cardiac magnetic resonance images using a capsule network. J X-Ray Sci Technol. 2020;28:541–53.
[24] Benjamini D, Iacono D, Komlosh ME, Perl DP, Brody DL, Basser PJ. Diffuse axonal injury has a characteristic multidimensional MRI signature in the human brain. Brain. 2021;144:800–16.
[25] Yu XF, Yin WW, Huang CJ, et al. Risk factors for relapse and nomogram for relapse probability prediction in patients with minor ischemic stroke. World J Clin Cases. 2021;9:9440–51.
[26] Hilbert A, Ramos LA, van Os HJA, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput Biol Med. 2019;115:103516.
[27] Kim BJ, Lee KJ, Park EL, et al. Prediction of recurrent stroke among ischemic stroke patients with atrial fibrillation: development and validation of a risk score model. PLoS One. 2021;16:e0258377.
[28] Wang K, Shou Q, Ma SJ, et al. Deep learning detection of penumbral tissue on arterial spin labeling in stroke. Stroke. 2020;51:489–97.
[29] Wang Y, Jing J, Meng X, et al. The Third China National Stroke Registry (CNSR-III) for patients with acute ischaemic stroke or transient ischaemic attack: design, rationale and baseline patient characteristics. Stroke Vasc Neurol. 2019;4:158–64.
[30] Lee H, Lee EJ, Ham S, et al. Machine learning approach to identify stroke within 4.5 hours. Stroke. 2020;51:860–6.
[31] Wang J, Zhang J, Gong X, Zhang W, Zhou Y, Lou M. Prediction of large vessel occlusion for ischaemic stroke by using the machine learning model random forests. Stroke Vasc Neurol. 2022;7:94–100.
[32] Zhou Y, Yang T, Gong Y, et al. Pre-hospital delay after acute ischemic stroke in central Urban China: prevalence and risk factors. Mol Neurobiol. 2017;54:3007–16.
[33] Zhang X, Lv W, Xu J, et al. The contribution of inflammation to stroke recurrence attenuates at low LDL-C levels. J Atheroscler Thromb. 2022;29:1634–45.
[34] Chen J, Li S, Zheng K, et al. Impact of smoking status on stroke recurrence. J Am Heart Assoc. 2019;8:e011696.
[35] Xu J, Zhang X, Jin A, et al. Trends and risk factors associated with stroke Recurrence in China, 2007-2018. JAMA Netw Open. 2022;5:e2216341.
[36] Xu YY, Chen WQ, Wang MX, et al. Lipid management in ischaemic stroke or transient ischaemic attack in China: result from China National Stroke Registry III. BMJ Open. 2023;13:e069465.
[37] Pan Y, Wangqin R, Li H, et al. LDL-C levels, lipid-lowering treatment and recurrent stroke in minor ischaemic stroke or TIA. Stroke Vasc Neurol. 2022;7:276–84.
[38] Wang Y, Fan H, Duan W, et al. Elevated stress hyperglycemia and the presence of intracranial artery stenosis increase the risk of recurrent stroke. Front Endocrinol (Lausanne). 2022;13:954916.
[39] Chung JY, Lee BN, Kim YS, Shin BS, Kang HG. Sex differences and risk factors in recurrent ischemic stroke. Front Neurol. 2023;14:1028431.
[40] Bangad A, Abbasi M, de Havenon A. Secondary ischemic stroke prevention. Neurotherapeutics. 2023;20:721–31.
[41] Lu J, Hutchens R, Hung J, et al. Performance of multilabel machine learning models and risk stratification schemas for predicting stroke and bleeding risk in patients with non-valvular atrial fibrillation. Comput Biol Med. 2022;150:106126.