BMC Neurology

RESEARCH Open Access

Development of machine learning-based models for predicting risk factors in acute cerebral infarction patients: a clinical retrospective study

Changqing Yang[1,2†], Renlin Hu[3†], Shilan Xiong[4,5], Zhou Hong[3], Jiaqi Liu[6], Zhuqing Mao[7*] and Mingzhu Chen[4,5*]

Abstract

Objectives The aim of this study was to develop machine learning-based models for predicting acute cerebral infarction (ACI) in patients.

Methods We extracted the data of ACI patients and non-ACI patients (as control) from two hospitals. The Lasso algorithm was employed to select the most crucial features associated with ACI. Five machine learning algorithms-based models *were trained, which was performed with 10-fold cross-validation*. Then, the area under the receiver operating characteristic curve (AUC), accuracy, and F1-score were *calculated in the training models*. *Accordingly, the training models with excellent performance was selected as the final predictive model*. The relative importance of variables was analyzed and ranked.

Results A total of 150 patients were diagnosed with ACI (50.00%), with a higher proportion of males (70.67% vs. 44.00%) compared to the non-ACI patients. The logistic regression model exhibited a good performance in predicting ACI *in the training set*, as evidenced by its highest AUC, accuracy, sensitivity, and F1-score. Furthermore, feature importance analysis showed that blood glucose, gender, smoking history, serum homocysteine, folic acid, and C-reactive protein were the top six crucial variables of the logistic regression.

Conclusions In our work, the ACI risk prediction model developed by the logistic regression exhibited excellent performance. This could contribute to the identification of risk variables for ACI patients and enables clinicians timely and effective interventions.

Keywords Acute cerebral infarction, Machine learning, Prediction model, Risk factors

[†]Changqing Yang and Renlin Hu contributed equally to this work.

*Correspondence:
Zhuqing Mao
18846146118@163.com
Mingzhu Chen
mingzhuchen2013@163.com
[1]Department of Hematology, Affiliated Hospital 6 of Nantong University, 02 Xinduxi Road, Yancheng 224000, China
[2]Department of Hematology, Yancheng Third People's Hospital, 02 Xinduxi Road, Yancheng 224000, China
[3]Department of Internal Medicine Neurology, Wuhan Fifth Hospital, 122 Xianzheng Street, Wuhan 430050, China
[4]Department of Neurology, Affiliated Hospital 6 of Nantong University, 02 Xinduxi Road, Yancheng 224000, China
[5]Department of Neurology, Yancheng Third People's Hospital, 02 Xinduxi Road, Yancheng 224000, China
[6]School of Medicine of Nantong University, 19 Qixiu Road, Nantong 226000, China
[7]Department of Neurology, Fushun Central Hospital, 05 Xincheng Road, Jinzhou 113000, China

Introduction

Acute cerebral infarction (ACI) is a cerebrovascular disease characterized by sudden occlusion of cerebral vessels, which results in necrosis of brain tissue and a neurological deficit [1]. It has a high incidence, disability rate, and mortality rate, emerging as a significant threat to human health [2]. Cerebral thrombosis and cerebral embolism are common clinical types of ACI that manifest with abrupt collapse, unconsciousness, speech disorder, sensory, or motor abnormalities. ACI profoundly impacts the quality of life and imposes an enormous economic burden on patients, families, and society. Therefore, rapid and accurate diagnosis, along with timely and effective treatment, is crucial for ACI patients. The imaging methods remain the preferred means of detecting ACI in patients, such as computed tomography (CT) and magnetic resonance imaging (MRI). However, these detection methods lack sensitivity in early ACI detection and cannot predict embolus formation. Consequently, clinical studies on risk prediction models of ACI are still in progress, and it is important for human health to explore reliable models to predict the occurrence of ACI [3, 4].

With advancements in detection technology, expansion of detection parameters, and widespread adoption of laboratory management system, a substantial volume of laboratory data is generated during patient's hospital visits. However, clinicians typically prioritize the identification of marked abnormal parameters while overlooking a considerable amount of other test data and the interconnectedness between laboratory parameters, leading to an underestimation of the true potential effect of these data [3]. The search for highly sensitive laboratory indicators to predict ACI and guide subsequent clinical decisions may emerge as a novel avenue of research. Previous studies have predominantly focused on employing traditional statistical models to analyze or predict risk factors associated with ACI [5, 6]. While these models offer some explanatory power in elucidating the correlation between laboratory parameters and ACI, they may still face challenges when dealing with complex clinical data and exhibit less accurate predictive performance.

Machine learning is an emerging discipline based on artificial intelligence and can automatically learn and deal with intricate interrelationships of data. Different from conventional regression models, machine learning effectively manages variable collinearity by regularization to prevent overfitting. It proficiently establishes predictive models for muti-field and muti-factor events, particularly within the medical domain [7, 8]. Therefore, the objective of this study was to develop machine learning models for predicting ACI based on accumulated laboratory data. This will provide a valuable reference for clinical ACI risk assessment.

Methods

Study objective and data source

The data of the patients were obtained from the two hospitals, including Fushun Central Hospital and Yancheng Third People's Hospital. The protocol of this study was approved and supervised by the Ethics Review Committee of Fushun Central Hospital (ethics number: 2023013) and Yancheng Third People's Hospital (ethics number: 2024-13). This study complies with the Declaration of Helsinki (revised in 2013). According to the criteria issued by the Chinese Guidelines for the diagnosis and treatment of acute ischemic stroke (2018), patients were diagnosed with ACI using MRI at Fushun Central Hospital from November 2019 to November 2020. Inclusion criteria for ACI patients: (1) age ≥ 18 years; and (2) admission within 48 h of symptom onset. Exclusion criteria for ACI patients: (1) infectious disease within 2–4 weeks; (2) severe immune system diseases; (3) malignant tumors; (4) serious vital organ diseases, such as cardiopulmonary, liver, and kidney disease; (5) use of steroids or immunosuppressants; (6) cerebral hemorrhage; (7) use of anti-inflammatory drugs; (8) hematological diseases; and (9) inability to undergo MRI examination. Then, inclusion criteria for non-ACI patients: patients in Department of Neurology at Fushun Central Hospital or Yancheng Third People's Hospital at the same period, except ACI patients. Exclusion criteria for non-ACI patients: (1) a history of ACI; (2) severe immune system diseases; (3) severe infections; and (4) malignant tumors. Due to the retrospective nature of this study, patient's informed consent was waived by the Ethics Review Committee of the hospitals.

Data collection and preprocessing of data

Clinical characteristics information was collected by the electronic medical record (EMR) system of the hospitals. Clinical characteristics of the patients included age, gender, body surface area (BSA), body mass index (BMI), and smoking history. Hypertension and diabetes information of patients were recorded. Additionally, laboratory indicators were tested within 48 h of admission, mainly including serum total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), fasting blood glucose, blood urea nitrogen (BUN), serum creatinine (sCr), C-reactive protein (CRP), brain natriuretic peptide (BNP), cardiac troponin T (cTnT), fibrinogen (Fib), serum sodium (Na^+), serum potassium (K^+), folic acid, homocysteine (Hcy), lithic acid, and vitamin B_{12} ($VitB_{12}$). Routine blood tests were conducted to collect counting or percentage information of white blood cells (WBC), leukomonocytes, neutrophils, monocytes, eosinophiles, and basophiles. All data in this study were collected through manual review of the medical records. Among all the variables, the overall rate of missing data was 0.194%. Missing data were supplemented using multiple imputation method *by the R programming language, specifically utilizing the "mice" package (version 3.12.0). In this process, the 10 imputed datasets were generated, and the average data was selected as the final data to be included in the model training.* Consequently, missing data on CRP were supplemented. Additionally, no abnormal data values were found in this study (Table S1 and Table S2), and all features were not normalized, standardized, or processed through other manner. Four categorical variables (including gender, smoking history, hypertension, diabetes) were binary. Gender was coded as "male" = 0 and "female" = 1. "Smoking history," "hypertension," and "diabetes" were coded as 1 if present, 0 if absent. There are not any features that were transformed or encoded.

Statistical analysis

Continuous variables were expressed as mean ± standard deviation or median (interquartile range). The differences in continuous variables between two groups were analyzed using Student's t-test or Mann-Whitney U test. Categorical variables were shown as number (percentage) and analyzed using the Chi-squared test. All statistical analyses were performed using R version 4.2.3 and python version 3.11.4, and $p < 0.05$ represents significant difference.

Model training and cross-validation

The Lasso algorithm was employed to select critical features, which could avoid overfitting and find optimal parameters. Subsequently, the selected variables in this study were used to train machine learning models, which *was performed with* 10-fold cross-validation. *In this process, according to the model training and cross-validation process, the data were randomly divided into a training subset or a validation subset according to a ratio of 9:1.* Meanwhile, the experiment's repeatability was ensured by employing fixed random seeds during the partitioning of the data sets. This approach can effectively guarantee consistent division results between the training and validation *subsets* when each time the program is executed. The training subset (90%) was used for model development, while the validation subset (10%) was employed *for hyperparameter tuning. In this study, five* machine learning-based algorithms, including logistic regression, LightGBM, complement Naive Bayes (CNB), support vector machine (SVM), and multi-layer perceptron neural network (MLP), were employed for the model prediction of ACI risk in patients.

Logistic regression model

The logistic regression model employed the liblinear solver due to its high efficiency with small datasets. L2 regularization was utilized to prevent overfitting and control the complexity of the model. The regularization parameter (C) was tuned using grid search with cross-validation to determine the optimal value. Feature scaling was performed to ensure model convergence and stability.

LightGBM model

For the LightGBM model, the Gradient Boosting Decision Tree (gbdt) boosting type was utilized. The tree parameters included a maximum depth of 20, a maximum tree count of 5, a maximum leaf count of 5, and a learning rate of 0.001. Bayesian optimization was employed for hyperparameter tuning, allowing efficient exploration of the parameter space.

CNB model

The CNB model implemented the complement version of the Naive Bayes classifier to handle class imbalance more effectively. The alpha parameter, controlling additive smoothing, was adjusted and tuned using cross-validation. The implementation was based on the scikit-library's Complement NB class, ensuring a robust and standardized approach.

SVM model

For the SVM model, a radial basis function (RBF) kernel was used, suitable for non-linear data. The regularization parameter (C) and the kernel coefficient (gamma) were both tuned using a grid search with cross-validation. Specifically, values for C ranged from 0.1 to 10, and for gamma from 0.01 to 1, ensuring optimal hyperparameters were identified.

MLP model

The MLP model architecture consisted of three hidden layers with 128, 64, and 32 neurons, respectively. The ReLU (Rectified Linear Unit) was used as the activation function for the hidden layers, and a softmax activation function was employed for the output layer. Categorical cross-entropy was used as the loss function, appropriate for the classification task. The model was optimized using the Adam optimizer with an initial learning rate of 0.001. Early stopping was employed with a patience of 10 epochs to prevent overfitting, and the learning rate was adjusted using a scheduler based on cross-validation loss improvement.

Model selection

The area under the receiver operating curve (AUC), accuracy (ACC), sensitivity, specificity, F1-score, positive predictive value (PPV), and negative predictive value (NPV) were calculated *in the training models.* Generally, AUC value of machine learning-based model greater than 0.8

indicates an excellent predictive ability. The Delong test was used for the significance test of AUCs among models, and $p<0.05$ represents a significant difference. Accordingly, these indicators were comprehensively used in our study to determine the final predictive model. Eventually, the crucial features selected by the Lasso algorithm were input into the final predictive model, and then, feature importance was analyzed to obtain the weight of variables affecting ACI. Figure 1 shows the flowchart of this study.

Results

Characteristics baseline

Patients were categorized into two groups based on MRI detection: the ACI group ($n=150$, 50%) and the non-ACI group ($n=150$, 50%). In the ACI group, the median age was 66 years old, whereas in the non-ACI group it was 63 years old. Regarding gender distribution, there were 106 males (70.67%) in the ACI group and 66 males (44.00%) in the non-ACI group. Analysis of baseline characteristics revealed significant differences between the two groups with regards to gender, smoking history, BMI, hypertension, diabetes, TC, LDL-C, HDL-C, blood glucose, Serum Na^+, cTnT, Hcy, folic acid, $VitB_{12}$, Fib, CRP, WBC, percentage of neutrophils, percentage of leukomonocytes, percentage of monocytes, percentage of eosinophiles, percentage of basophiles, neutrophils count, basophiles count, and NLR (*p<0.05*, Table 1). Furthermore, no differences were observed between the two groups in terms of age, BSA, TG, BUN, sCr, BNP, lithic acid, leukomonocytes count, or monocytes count (Details are shown in Table 1).

Feature selection of Lasso algorithm

The Lasso algorithm was conducted for candidates' selection, resulting in the identification of 28 features out of 36 variables in this study (1.28:1 ratio, Fig. 2). The selected variables were CRP, BNP, neutrophils, basophiles, eosinophiles, leukomonocytes, percentage of basophiles, percentage of monocytes, percentage of eosinophiles, WBC, Fib, VitB12, folic acid, Hcy, lithic acid, serum Na^+, serum K^+, BUN, blood glucose, HDL-C, TG, TC, hypertension, BSA, BMI, smoking history, gender, and age. Subsequently, the variables were incorporated into machine learning-based models for risk prediction.

Selection of machine learning models

Compared to other models, the logistic regression model had the highest AUC *in the training data (0.913*, Fig. 3*)*. This was further confirmed by using Delong test (Table S3 and Table S4). The logistic regression model showed the greatest ACC (0.833), sensitivity (0.899), PPV (0.837),

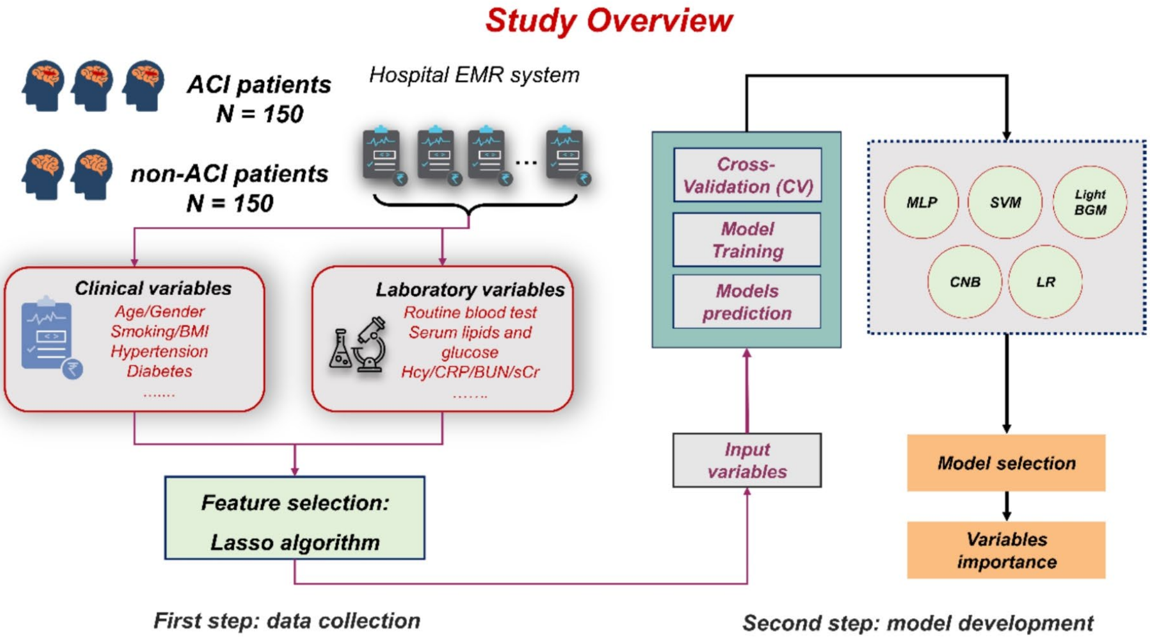

Fig. 1 The flowchart of this study. ACI, acute cerebral infarction; SVM, support vector machine; MLP, multi-layer perceptron neural network; CNB, complement Naive Bayes; LR, logistic regression; EMR, electronic medical record. The ACI group ($N=150$), the non-ACI group ($N=150$)

Table 1 Baseline characteristics of patients in the non-ACI group and the ACI group

Variables		Total (*N* = 300)	Non-ACI group (*N* = 150)	ACI group (*N* = 150)	Statistic values	*p*-value
Age (year)		65.0 (57.0, 71.0)	63.0 (56.0, 70.0)	66.0 (59.0, 72.0)	-1.805	0.071
Gender (%)	**Male**	172 (57.333)	66 (44.000)	106 (70.667)	21.802	**<0.001**
	Female	128 (42.667)	84 (56.000)	44 (29.333)		
Smoking history (%)	**No**	193 (64.333)	127 (84.667)	66 (44.000)	54.055	**<0.001**
	Yes	107 (35.667)	23 (15.333)	84 (56.000)		
BMI (kg/m^2)		24.22 (22.32, 26.56)	24.77 (22.66, 27.34)	24.21 (22.13, 25.95)	2.245	**0.025**
BSA (m^2)		1.72 (1.60, 1.81)	1.68 (1.57, 1.81)	1.75 (1.65, 1.82)	-1.832	0.067
Hypertension (%)	**No**	147 (49.000)	82 (54.667)	65 (43.333)	3.855	**0.050**
	Yes	153 (51.000)	68 (45.333)	85 (56.667)		
Diabetes (%)	**No**	228 (76.000)	134 (89.333)	94 (62.667)	29.240	**<0.001**
	Yes	72 (24.000)	16 (10.667)	56 (37.333)		
Serum TC (mmol/L)		4.75 (4.10, 5.31)	4.39 (3.82, 5.08)	4.98 (4.33, 5.62)	-4.703	**<0.001**
Serum TG (mmol/L)		1.48 (1.06, 1.95)	1.42 (1.04, 1.88)	1.52 (1.11, 2.03)	-1.183	0.237
Serum LDL-C (mmol/L)		2.58 (2.14, 3.00)	2.49 (1.92, 3.00)	2.64 (2.26, 3.05)	-2.027	**0.043**
Serum HDL-C (mmol/L)		1.28 (1.08, 1.50)	1.25 (1.05, 1.46)	1.32 (1.10, 1.57)	-2.485	**0.013**
Blood glucose (mmol/L)		6.50 (5.10, 9.00)	5.16 (4.53, 6.80)	8.20 (6.20, 11.90)	-9.786	**<0.001**
BUN (mmol/L)		5.74 (4.73, 6.98)	5.58 (4.72, 6.86)	5.86 (4.73, 7.10)	-1.208	0.227
sCr (μmol/L)		65.30 (54.60, 78.00)	63.80 (53.90, 76.10)	67.20 (55.30, 80.90)	-1.487	0.137
Serum Na$^+$(mmol/L)		139.0 (137.0, 141.1)	140.8 (138.2, 142.5)	138.0 (136.0, 139.0)	7.823	**<0.001**
Serum K$^+$(mmol/L)		3.93 (3.74, 4.15)	3.97 (3.74, 4.17)	3.90 (3.73, 4.15)	1.042	0.298
BNP (pg/mL)		151.29 (79.47, 222.08)	151.29 (108.01, 208.44)	135.10 (63.00, 302.80)	0.454	0.650
cTnT (pg/mL)		0.009 (0.007, 0.012)	0.009 (0.007, 0.012)	0.008 (0.006, 0.012)	2.201	**0.027**
Lithic acid (μmol/L)		310.0 (244.0, 372.0)	309.8 (242.8, 362.0)	308.0 (244.0, 382.0)	0.043	0.967
Hcy (μmol/L)		11.4 (8.9, 15.0)	11.1 (8.5, 13.7)	11.6 (9.1, 17.4)	-2.324	**0.020**
Folic acid (ng/mL)		8.30 (5.78, 11.11)	9.97 (6.84, 12.49)	7.28 (4.47, 9.49)	5.551	**<0.001**
VitB$_{12}$(ng/mL)		350.6 (284.0, 417.0)	340.0 (242.0, 417.9)	357.2 (316.4, 414.8)	-2.773	**0.006**
Fib (pg/L)		2.88 (2.51, 3.31)	2.77 (2.38, 3.27)	2.98 (2.56, 3.40)	-2.578	**0.010**
CRP (mg/L)		1.40 (0.85, 3.32)	1.18 (0.50, 1.68)	2.40 (1.25, 5.37)	-7.264	**<0.001**
WBC (10^9/L)		6.60 (5.62, 8.26)	6.32 (5.16, 7.86)	7.01 (5.94, 8.63)	-3.754	**<0.001**
Percentage of neutrophils (%)		66.85 ± 11.46	63.79 ± 11.79	69.91 ± 10.24	-4.786	**<0.001**
Percentage of leukomonocytes (%)		24.86 ± 9.31	26.85 ± 9.28	22.86 ± 8.91	3.788	**<0.001**
Percentage of monocytes (%)		5.80 (4.60, 7.20)	6.60 (5.30, 8.40)	5.10 (4.40, 6.60)	5.195	**<0.001**
Percentage of eosinophiles (%)		1.10 (0.50, 2.20)	1.30 (0.60, 2.60)	1.0 (0.50, 1.80)	2.120	**0.034**
Percentage of basophiles (%)		0.30 (0.10, 0.40)	0.30 (0.20, 0.50)	0.20 (0.10, 0.30)	3.990	**<0.001**
Neutrophils (10^9/L)		4.42 (3.45, 5.88)	3.79 (3.02, 5.42)	4.80 (4.04, 6.20)	-5.033	**<0.001**
Leukomonocytes (10^9/L)		1.59 (1.20, 2.03)	1.60 (1.21, 2.11)	1.49 (1.180, 2.0)	0.895	0.371
Monocytes (10^9/L)		0.39 (0.31, 0.50)	0.39 (0.32, 0.52)	0.38 (0.29, 0.47)	1.320	0.187
Eosinophiles (10^9/L)		0.08 (0.04, 0.15)	0.09 (0.04, 0.17)	0.07 (0.04, 0.12)	1.890	0.059
Basophiles (10^9/L)		0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.01 (0.01, 0.03)	3.079	**0.002**
NLR		2.70 (1.91, 3.82)	2.36 (1.71, 3.51)	3.06 (2.20, 4.61)	-4.185	**<0.001**

Continuous variables were presented as mean ± standard deviations (SD) or median (interquartile spacing). Categorical variables were presented as numerical values and proportions. $p < 0.05$ represents a significant statistical difference

ACI, acute cerebral infarction; BMI, body mass index; BSA, body surface area; TC, total cholesterol; TG, triglyceride; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; BUN, blood urea nitrogen; sCr, serum creatinine; BNP, brain natriuretic peptide; cTnT, cardiac troponin T; Hcy, homocysteine; Fib, fibrinogen; CRP, C-reactive protein; WBC, white blood cell; NLR, neutrophils-to-leukomonocyte ratio

NPV (0.845), and F1-score (0.837) *in the training data (*Table 2*). Furthermore*, LightGBM also performed well but was slightly inferior to the logistic regression in terms of AUC and ACC. Comprehensively considering these parameters of *the training models*, the logistic regression model was finally selected as the predictive model in this study.

Relative importance of variables in logistic regression model

The relative importance of variables in the logistic regression model for predicting ACI risk is shown in Fig. 4. There are general evidence trends: while the weight of these variables varies, blood glucose, gender, smoking history, Hcy, folic acid, CRP, TC, and age are more critical than other risk factors. The importance of high-ranking

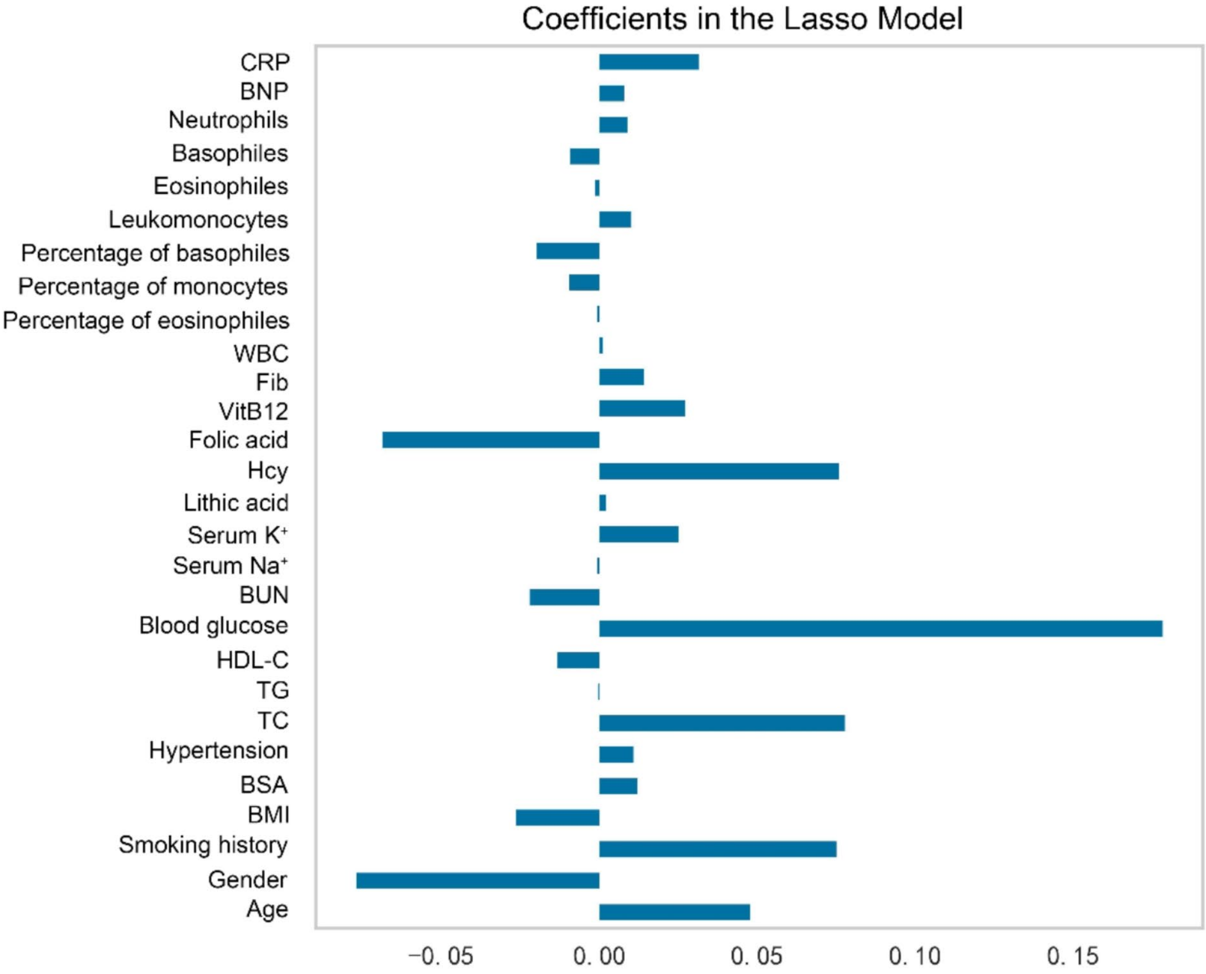

Fig. 2 Clinical features selection using the Lasso algorithm. 28 out of 36 clinical variables were finally selected *to train machine learning models*

variables in the logistic regression model is arranged in descending order (the top six): blood glucose, gender, smoking history, Hcy, folic acid, and CRP.

Discussion

The popularity of the hospital's EMR system has facilitated the storage of a vast amount of patient information in the hospital databases, including demographics, chronic diseases, laboratory tests, and imaging data. This wealth of the data presents novel opportunities for data-driven artificial intelligence models to accurately predict the occurrence, development, and prognosis of diseases [9–11]. Darabi et al. successfully developed fifteen machine learning-based models that effectively predicted risk factors associated with 30-day readmission following stroke using sixty-one clinical variables [12]. Similarly, Li et al. have achieved success by developing machine learning-based prediction models for cerebral infarction primarily utilizing biochemical variables [3]. These studies collectively demonstrated robust predictive performance of machine learning-based models. In this study, based on the clinical variables data collected within 48 h after admission, *we developed five risk prediction models* for ACI using machine learning algorithms. *Among all training models*, the logistic regression model demonstrated superior predictive performance, *which was* attributed by its highest AUC, ACC, sensitivity, and F1-score. *Furthermore, this was confirmed by Delong test. Accordingly, the logistic regression model was considered as the final predictive model in this study.*

Although CT and MRI detection have their own unique advantages in the diagnosis of ACI, they are unable to predict embolism formation or assess the probability of ACI in patients. Previous studies have utilized machine learning method to reveal hidden relationship between routine biochemical tests and diseases, such as cardiovascular and cerebrovascular diseases, yielding valuable medical insights that enable test results to provide information for disease diagnosis beyond the reference range [8, 13]. Similarly, the clinical data of 300

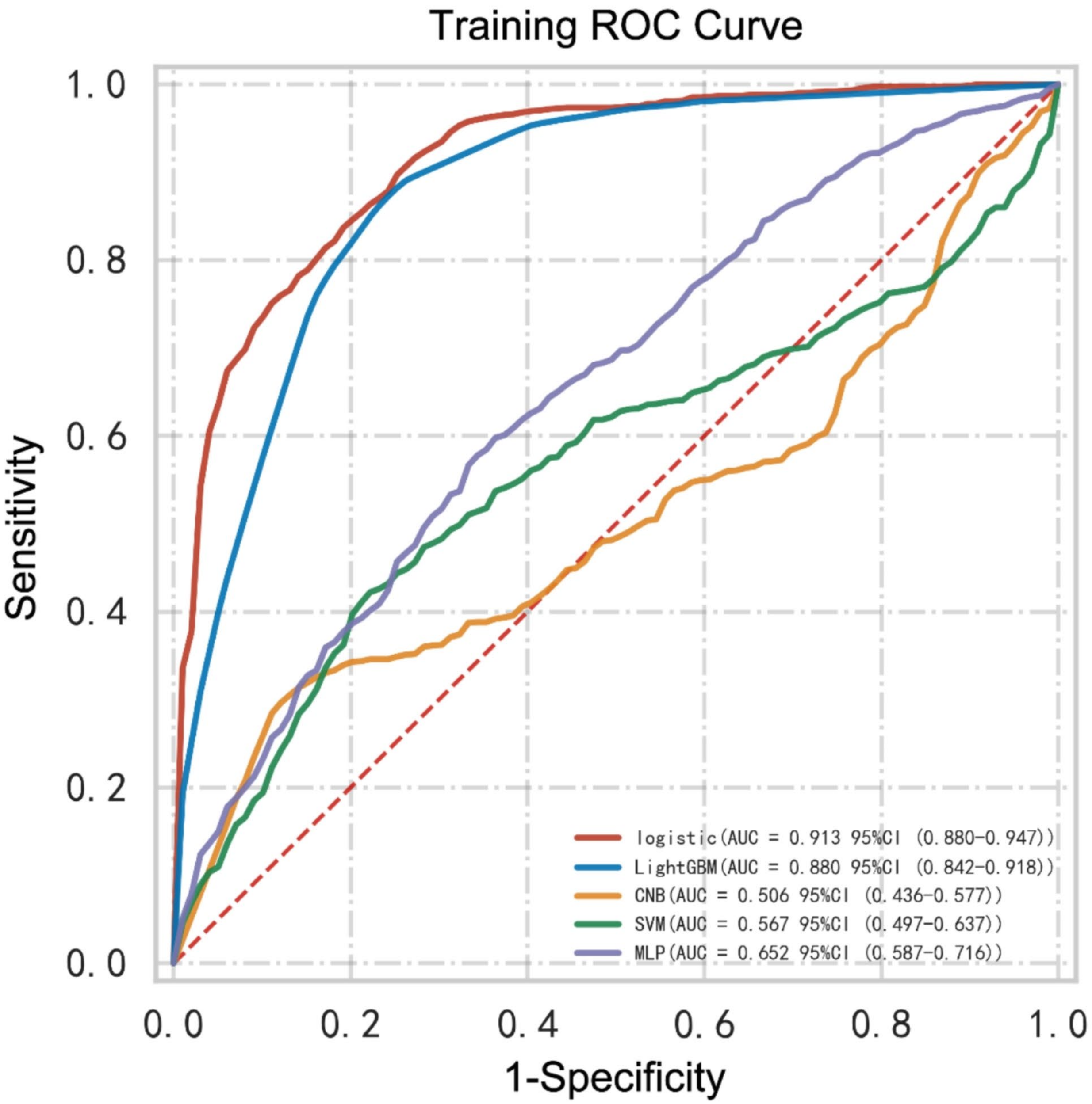

Fig. 3 The receiver operating characteristic curves (ROCs) of machine learning training models. AUC, area under the ROC; SVM, support vector machine; MLP, multi-layer perceptron neural network; CNB, complement Naive Bayes

patients (including 150 ACI patients and 150 non-ACI patients) were included in this study. Initially, the Lasso algorithm was employed to select the variables closely related to ACI occurrence. *Subsequently, these selected variables were used for model training (including 10-fold cross-validation process).* Ultimately, blood glucose, gender, smoking history, Hcy, folic acid, and CRP emerged as the top six characteristics in the logistic regression model that exerted significant influence on predicting ACI. This finding provides clinicians with a risk model capable of accurately identifying ACI cases and may have potential clinical significance.

Metabolic syndrome is considered a significant risk factor for the development and progression of ACI. Previous studies have demonstrated a notable impact of elevated fasting blood glucose on ACI in patients [3, 14], which can be attributed to impaired cerebrovascular endothelial cell function caused by high blood glucose levels. Furthermore, high-level inflammation induced by high blood glucose involves in the development of ACI.

Table 2 Comparison of machine learning *training models*

Models	AUC	ACC	Sensitivity	Specificity	PPV	NPV	F1-score
Logistic Regression	0.913 (0.880–0.947)	0.833 (0.825–0.841)	0.899 (0.860–0.937)	0.828 (0.780–0.877)	0.837 (0.805–0.869)	0.845 (0.810–0.879)	0.837 (0.826–0.848)
LightGBM	0.880 (0.842–0.918)	0.788 (0.725–0.851)	0.846 (0.797–0.894)	0.759 (0.710–0.808)	NA	0.774 (0.709–0.839)	NA
CNB	0.506 (0.436–0.577)	0.59 (0.586–0.595)	0.306 (0.299–0.312)	0.882 (0.873–0.891)	0.718 (0.703–0.732)	0.557 (0.554–0.560)	0.429 (0.422–0.436)
SVM	0.567 (0.497–0.637)	0.609 (0.603–0.616)	0.459 (0.422–0.496)	0.767 (0.727–0.806)	0.664 (0.644–0.684)	0.584 (0.577–0.590)	0.54 (0.520–0.560)
MLP	0.652 (0.587–0.716)	0.63 (0.595–0.664)	0.579 (0.477–0.680)	0.688 (0.628–0.748)	0.646 (0.615–0.678)	0.628 (0.582–0.674)	0.6 (0.534–0.666)

All values are shown as mean (95% confidence interval)

AUC, area under the receiver operating characteristic curve; ACC, accuracy; PPV, positive prediction value; NPV, negative prediction value; MLP, multi-layer perceptron neural network; SVM, support vector machine; CNB, complement Naive Bayes. NA: Not applicable

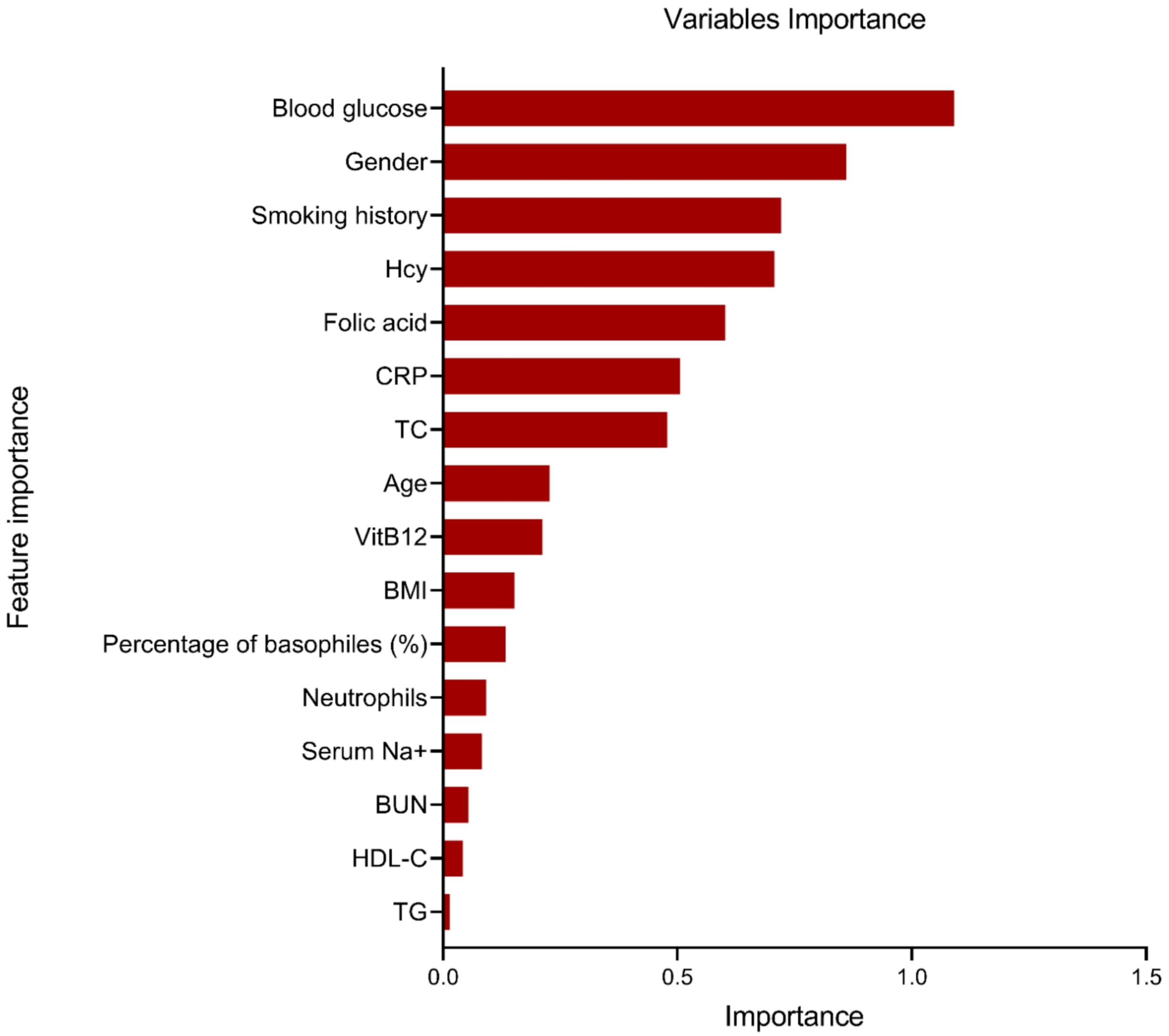

Fig. 4 The relative importance of the variables in the logistic regression model is in decreasing order. The relative importance of high-ranking variables in the logistic regression model is arranged in descending order (the top six): blood glucose, gender, smoking history, Hcy, folic acid, and CRP

A multi-center observational cohort study reported the association of baseline fasting plasma glucose with 1-year mortality in non-diabetic patients with ACI, which highlights the significance of fast blood glucose on ACI [15]. Our study revealed that the ACI group had significantly higher blood glucose levels compared to the non-ACI group (8.20 vs. 5.16 mmol/L), suggesting abnormal glucose metabolism and confirming previous research findings in ACI patients. Importantly, blood glucose was identified as the most crucial risk factor for ACI.

Gender has been reported to be associated with the occurrence and progression of cardiovascular and cerebrovascular diseases, including acute or chronic cerebral infarction [16, 17]. Males exhibit a higher risk of stroke, vascular cognitive impairment, and dementia compared to females during lifetime [17]. This disparity may be attributed to the direct impact of androgen on the cerebrovascular system, simultaneously influenced by various factors, such as age, hormone levels, and disease status [17]. The study has found that females between the ages of 45 to 74 years old have a significant lower likelihood of experiencing cerebral infarction than males due to a potential protective effect conferred by estrogen [18]. In this study, the proportion of male patients in the ACI group (70.67%) was significantly higher than that in the non-ACI group (44.00%). Our results revealed an existed difference in the proportion of males, which was related to the development of ACI. Furthermore, gender was considered as a secondary influencing factor to blood glucose in the logistic regression model. A clinical case-control study (the INTERSTROKE study) was conducted in 22 countries, and the results showed that smoking was all significant factor for ischemic stroke, which may be related to inflammation and vascular injury [19]. In this study, the ACI group had more patients with smoking compared to the non-ACI group (56.00% vs. 15.33%). Given the weight effect of smoking history in the logistic regression model, the smoking should be on the alert of clinicians and patients.

Hcy, the main metabolite of methionine intermediates, is a reactive substance that causes vascular injury. Elevated serum levels of Hcy are associated with cerebral infarction, which may be attributed to inflammatory response and damage inflicted on vascular endothelium [11, 20]. Our study identified serum Hcy as a crucial biochemical indicator for predicting ACI. Folic acid is an important determinant of Hcy metabolism and thereby participates indirectly in the progression of ACI. Previous study has demonstrated the negative relationship of folic acid levels and the risk of ACI [21]. Increasing levels of folic acid can reduce serum Hcy concentrations, improve the degree of carotid atherosclerosis, and protect vascular endothelium in patients [22]. Our study found a markedly decreased folic acid levels in the ACI group than the non-ACI group (7.28 vs. 9.97 ng/mL). This prevents a clearance of Hcy from the body, thereby with an elevated Hcy levels in ACI patients. Long-term exposure to inflammatory factors can trigger vascular inflammation and even ischemic brain injury [23, 24]. CRP, a non-specific inflammatory factor synthesized by the liver, is elevated in response to trauma, inflammation, or stress. Clinical study found that CRP was a sensitive predictor of stroke and cerebrovascular events in the general population [25]. In this study, serum CRP levels were obviously higher in the ACI group than the non-ACI group (2.40 vs. 1.18 mg/L), highlighting its importance as a predictive factor of ACI. More importantly, five out of the top six predictors in ACI models were associated with either vascular injury or vascular inflammation. Hence, serum biomarkers related to vascular injury or inflammation deserve more attention in future clinical practice.

It must be noticed that our study has the following advantages. Firstly, few studies have explored the risk factors associated with ACI using machine learning-based algorithms and routine biochemical parameters. Our study could construct machine learning models to predict ACI risk using biochemical data collected within short-time, which exhibited the innovations of the study and enriched the existing literature. Furthermore, our study might be helpful for clinicians to early identify risk variables of ACI patients. Secondly, five machine learning-based models were used to predict ACI, among which the logistic regression model performed excellently, which to some extent, supports the accuracy and reliability of the result. Finally, considering numerous included variables in our dataset, these machine learning techniques could include higher-order interactions among variables. Our created model has analyzed a diverse list of potential predictors and identified the importance of the variables. However, the present study still has the following limitations. Firstly, this study is a retrospective study, which may lead to a certain degree of bias in the results. *Secondly, we used the average of multiple imputed datasets as the final data for model training when handing missing data. Indeed, multiple imputation is designed to account for the uncertainty and variability introduced by missing data. Averaging the imputed datasets can obscure this variability and lead to biased estimates. Thirdly, in this study, we defined all the data as a training set and did not divide it into a training set and a test set. The lack of test set results in inadequate model validation, a degree of difficulty in evaluating model generalization, inability to detect overfits, and inability to objectively compare the performance of different models. Fourthly, the sample size of the patients included in the models was relatively small, and there was a lack of external data validation, which may affect the universality of the model to some extent. These problems can affect the* *model training, validation, evaluation, and generalization to some extent and are expected to be addressed.* Finally, our research employed tree-based algorithms, such as decision trees, which are generally robust to feature scaling. Therefore, we initially decided not to perform feature scaling before model training. However, it is highly important to acknowledge that the absence of feature scaling may have some impact on our research findings. Although tree-based algorithms are insensitive to feature scaling, other non-tree algorithms like SVM, K-nearest neighbors (KNN), and neural networks can suffer from performance issues without proper feature scaling. Neglecting feature scaling can result in misinterpretation of the significance of features with different dimensions. Larger dimension features might be mistakenly perceived as more important, thereby affecting the interpretation of our model results. Additionally, incorporating feature scaling enhances stability and convergence during the training process; hence its omission may lead to longer training times. Accordingly, we must recognize potential effects arising from the lack of feature scaling in this study. Importantly, despite the absence of this process in this study, future studies should explore the specific impacts of feature scaling on different algorithms and implementations for a comprehensive understanding of its role and importance across various machine learning models. *Collectively, some* large prospective cohort studies (*including external data*) are urgently needed to validate the results of this study in the future.

Conclusion

Machine learning-based models were *trained and cross-validated* to predict risk factors of ACI in patients using routine biochemical parameters. The logistic regression model showed good predictive performance. This may help clinicians to identify high-risk patients early through this model and carry out timely and effective intervention treatment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12883-024-03818-6.

Supplementary Material 1

Acknowledgements

This work was supported by Extreme Smart Analysis platform (https://www.xsmartanalysis.com/).

Author contributions

Changqing Yang and Renlin Hu taken part in the design of this study, collected the data, analyzed the data, and drafted the draft manuscript. Shilan Xiong collected and analyzed the data. Zhou Hong and Jiaqi Liu managed and supervised the data. Zhuqing Mao and Mingzhu Chen were responsible for the whole study, designed the study, and reviewed the manuscript.

Funding

No funds were received in support of this work.

Data availability

The data set or the analytical code supporting the conclusion of this article are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The protocol of this study was approved and supervised by the Ethics Review Committee of Fushun Central Hospital (ethics number: 2023013) and Yancheng Third People's Hospital (ethics number: 2024-13). This study complies with the Declaration of Helsinki (revised in 2013). Due to a retrospective nature of this study, patient's informed consent was waived by the Ethics Review Committee of the Hospitals.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 19 September 2023 / Accepted: 22 August 2024
Published online: 31 August 2024

References

1. Suzuki H, Kanamaru H, Kawakita F, Asada R, Fujimoto M, Shiba M. Cerebrovascular pathophysiology of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage. Histol Histopathol. 2021;36(2):143–58.
2. Zhang H, Yang G, Dong A. Prediction Model between Serum Vitamin D and Neurological Deficit in Cerebral Infarction Patients Based on Machine Learning. Computational and mathematical methods in medicine. 2022;2022:2914484.
3. Li X, Wang Y, Xu J. Development of a machine learning-based risk prediction model for cerebral infarction and comparison with nomogram model. J Affect Disord. 2022;314:341–8.
4. Nishi H, Oishi N, Ogawa H, Natsue K, Doi K, Kawakami O, et al. Predicting cerebral infarction in patients with atrial fibrillation using machine learning: the Fushimi AF registry. J Cereb Blood flow Metabolism: Official J Int Soc Cereb Blood Flow Metabolism. 2022;42(5):746–56.
5. Zhang X, Hu Y, Hong M, Guo T, Wei W, Song S. Plasma thrombomodulin, fibrinogen, and activity of tissue factor as risk factors for acute cerebral infarction. Am J Clin Pathol. 2007;128(2):287–92.
6. Dong XL, Xu SJ, Zhang L, Zhang XQ, Liu T, Gao QY, et al. Serum resistin levels may contribute to an increased risk of Acute Cerebral infarction. Mol Neurobiol. 2017;54(3):1919–26.
7. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216–9.
8. Deo RC. Machine learning in Medicine. Circulation. 2015;132(20):1920–30.
9. Vahidy FS, Donnelly JP, McCullough LD, Tyson JE, Miller CC, Boehme AK, et al. Nationwide estimates of 30-Day readmission in patients with ischemic stroke. Stroke. 2017;48(5):1386–8.
10. Lichtman JH, Leifheit-Limson EC, Jones SB, Wang Y, Goldstein LB. Preventable readmissions within 30 days of ischemic stroke among Medicare beneficiaries. Stroke. 2013;44(12):3429–35.
11. Lv J, Zhang M, Fu Y, Chen M, Chen B, Xu Z, et al. An interpretable machine learning approach for predicting 30-day readmission after stroke. Int J Med Informatics. 2023;174:105050.
12. Darabi N, Hosseinichimeh N, Noto A, Zand R, Abedi V. Machine learning-enabled 30-Day readmission model for stroke patients. Front Neurol. 2021;12:638267.
13. Zhuang H, Bao A, Tan Y, Wang H, Xie Q, Qiu M, et al. Application and prospect of artificial intelligence in digestive endoscopy. Expert Rev Gastroenterol Hepatol. 2022;16(1):21–31.
14. Lip GY, Blann AD, Farooqi IS, Zarifis J, Sagar G, Beevers DG. Sequential alterations in haemorheology, endothelial dysfunction, platelet activation and thrombogenesis in relation to prognosis following acute stroke: the West

Birmingham Stroke Project. Blood Coagulation Fibrinolysis: Int J Haemostasis Thromb. 2002;13(4):339–47.

15. Zhang D, Liu Z, Liu P, Zhang H, Guo W, Lu Q, et al. Association of baseline fasting plasma glucose with 1-year mortality in non-diabetic patients with acute cerebral infarction: a multicentre observational cohort study. BMJ Open. 2023;13(9):e069716.
16. Longpré-Poirier C, Dougoud J, Jacmin-Park S, Moussaoui F, Vilme J, Desjardins G, et al. Sex and gender and allostatic mechanisms of Cardiovascular Risk and Disease. Can J Cardiol. 2022;38(12):1812–27.
17. Abi-Ghanem C, Robison LS, Zuloaga KL. Androgens' effects on cerebrovascular function in health and disease. Biology sex Differences. 2020;11(1):35.
18. Raz L. Estrogen and cerebrovascular regulation in menopause. Mol Cell Endocrinol. 2014;389(1–2):22–30.
19. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet (London England). 2010;376(9735):112–23.
20. Fu HJ, Zhao LB, Xue JJ, Wu ZX, Huang YP, Liu W, et al. Elevated serum homocysteine (hcy) levels may contribute to the pathogenesis of cerebral infarction. J Mol Neuroscience: MN. 2015;56(3):553–61.
21. Sone J, Mori K, Inagaki T, Katsumata R, Takagi S, Yokoi S, et al. Clinicopathological features of adult-onset neuronal intranuclear inclusion disease. Brain. 2016;139(Pt 12):3170–86.
22. Li C, Bu X, Liu Y. Effect of folic acid combined with pravastatin on arteriosclerosis in elderly hypertensive patients with lacunar infarction. Medicine. 2021;100(28):e26540.
23. De Meyer SF, Denorme F, Langhauser F, Geuss E, Fluri F, Kleinschnitz C. Thromboinflammation in stroke brain damage. Stroke. 2016;47(4):1165–72.
24. Iadecola C, Anrather J. The immunology of stroke: from mechanisms to translation. Nat Med. 2011;17(7):796–808.
25. Park SW, Park SS, Kim EJ, Sung WS, Ha IH, Jung B. Sex differences in the association between self-rated health and high-sensitivity C-reactive protein levels in koreans: a cross-sectional study using data from the Korea National Health and Nutrition Examination Survey. Health Qual Life Outcomes. 2020;18(1):341.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.