European Journal of Medical Research

RESEARCH

Open Access

Predicting 3-month poor functional outcomes of acute ischemic stroke in young patients using machine learning

Lamia Mbarek[1,2], Siding Chen[1,2,3], Aoming Jin[1,2], Yuesong Pan[1,2], Xia Meng[1,2], Xiaomeng Yang[1], Zhe Xu[1,2], Yong Jiang[1,2,3,4*] and Yongjun Wang[1,2,3,5,6,7,8*]

Abstract

Background Prediction of short-term outcomes in young patients with acute ischemic stroke (AIS) may assist in making therapy decisions. Machine learning (ML) is increasingly used in healthcare due to its high accuracy. This study aims to use a ML-based predictive model for poor 3-month functional outcomes in young AIS patients and to compare the predictive performance of ML models with the logistic regression model.

Methods We enrolled AIS patients aged between 18 and 50 years from the Third Chinese National Stroke Registry (CNSR-III), collected between 2015 and 2018. A modified Rankin Scale (mRS) ≥ 3 was a poor functional outcome at 3 months. Four ML tree models were developed: The extreme Gradient Boosting (XGBoost), Light Gradient Boosted Machine (lightGBM), Random Forest (RF), and The Gradient Boosting Decision Trees (GBDT), compared with logistic regression. We assess the model performance based on both discrimination and calibration.

Results A total of 2268 young patients with a mean age of 44.3 ± 5.5 years were included. Among them, (9%) had poor functional outcomes. The mRS at admission, living alone conditions, and high National Institutes of Health Stroke Scale (NIHSS) at discharge remained independent predictors of poor 3-month outcomes. The best AUC in the test group was XGBoost (AUC = 0.801), followed by GBDT, RF, and lightGBM (AUCs of 0.795, 0, 794, and 0.792, respectively). The XGBoost, RF, and lightGBM models were significantly better than logistic regression ($P < 0.05$).

Conclusions ML outperformed logistic regression, where XGBoost the boost was the best model for predicting poor functional outcomes in young AIS patients. It is important to consider living alone conditions with high severity scores to improve stroke prognosis.

Keywords Acute ischemic stroke, Young patients, Poor functional outcomes, Machine learning

[†]Lamia Mbarek and Siding Chen are co-authors.

*Correspondence:
Yong Jiang
jiangyong@ncrcnd.org.cn
Yongjun Wang
yongjunwang@ncrcnd.org.cn
Full list of author information is available at the end of the article

Background

Stroke is the second leading cause of death worldwide and the leading cause of death and disability in China [1, 2]. Acute ischemic stroke (AIS) or transient ischemic attack (TIA) accounts for approximately 80% of all strokes. Stroke was once a disease of the elderly, although there is no universally agreed-upon definition for young adult patients, most research specifies this age group as individuals between 18 and 50 years, which is the definition used in our study [3, 4]. Recent studies suggest that 10% to 15% of all strokes occur in young adults between the ages of 18 and 50, resulting in approximately 2 million young adults worldwide having a stroke each year, with the incidence increasing over the past decade [5, 6]. Severe functional outcomes affect about 20–25% of stroke patients [7]. As certainty, the degree of disability/dependence after a stroke was measured using the modified Rankin Scale (mRS), which ranges from 0 to 6, with an mRS of 6 indicating death. Young patients with poor functional outcomes have a significant impact on health due to high medical costs and reduced work productivity [8]. Therefore, accurate prediction of functional outcomes after stroke will facilitate post-stroke management and improve the distribution of healthcare services.

In recent years, machine learning (ML) techniques have been increasingly used to solve a variety of research problems, including diagnostic and clinical research, such as stroke [9, 10]. ML is a subfield of artificial intelligence in which a computer extracts past information and uses it to predict new information. It can self-optimize by learning complex systems containing many variables and data [11]. Various algorithms have been used in previous studies, such as logistic regression, random forest classifier (RF) [12], support vector machine (SVM) [13], fully connected deep neural network (DNN) [9], Catboost [14] and extreme gradient boosting (XGBoost) [15] have been used to predict poor functional outcomes in general patients with AIS. However, the identifying factors that predict disability in young patients (under 50 years) are unclear.

In this study, four ML models (XGBoost, Light Gradient Boosting Machine (lightGBM), RF, and Gradient Boosting Decision Trees (GBDT)) were developed and compared with logistic regression to predict poor 3-month functional outcomes in young AIS patients. ML models are increasingly utilized in healthcare due to their ability to handle complex, non-linear relationships within the data and their potential for high accuracy [16]. XGBoost, a powerful and popular gradient-boosting algorithm, was chosen for its ability to handle missing data, scalability, and high predictive performance [17]. LightGBM, another gradient-boosting framework, was selected for its efficiency and speed in handling large datasets [18]. The RF, known for its robustness and ability to handle noisy data [19], was included for comparison. The GBDT model, which is similar to XGBoost but with some differences in the underlying algorithm [20], was also employed. We select logistic regression as a benchmark to illustrate the strengths and weaknesses of the more complex models utilized in our study, thereby providing deeper insights into their comparative advantages. The rationale for choosing these models in our study was based on their proven effectiveness in handling complex relationships, capturing interactions within the data, and providing accurate predictions. Each model was selected for its specific strengths, such as computational efficiency, scalability, robustness against overfitting, and the ability to capture complex patterns in the data. By comparing these models with logistic regression and evaluating discrimination based on the area under the receiver operating characteristic curve (AUC), our study aimed to determine the most effective approach for predicting poor functional outcomes in young AIS patients using our third China National Stroke Registry (CNSR-III).

Patients and methods

Study population

The CNSR-III is a large-scale prospective registry of acute ischemic cerebrovascular events in China, enrolled patients with AIS or TIA between August 2015 and March 2018. CNSR-III encompasses 201 sites distributed across 22 provinces and four municipalities across China. Specifically, 163 central teaching hospitals and 38 urban hospitals were selected based on their comprehensive assessment, adequate research personnel, relevant experience, and qualified equipment. The study design and methods of CNSRIII have been previously published [21]. We include young patients aged between 18 and 50 in this study. Asymptomatic patients with cerebral infarction who had no signs or symptoms or refused to participate in the registry were excluded.

Predictors and data processing

This study presented a comprehensive list of the 55 variables summarized in Supplementary Table 1, including demographic characteristics (sex, age, BMI, living condition, marital status, education level), thrombolytic therapy (alteplase), history of smoking, history of alcohol consumption, medical history (hypertension, diabetes, dyslipidemia, migraine, stroke, TIA, heart disease, arterial fibrillation), family history (hypertension, diabetes, dyslipidemia, stroke, heart disease, cancer), laboratory data, neurological severity such as the National Institutes of Health Stroke Scale (NIHSS) at admission and discharge, mRS score, the different etiology classified according to the

trial of ORG 10172 in acute stroke treatment (TOAST), and secondary prevention treatment at discharge. We had 24 variables with missing values, and the rates were below 5%. We used linear interpolation to impute missing values for continuous variables and mode imputation for categorical variables. We randomly divided the total dataset into a training set and a test set in an 80:20 ratio. Feature selection, parameter tuning, and model training were performed on the training set, while validation was conducted on the test set. We utilized the Sequential Forward Selection (SFS) combined with the K Nearest Neighbors (KNN) technique for feature selection.

Patient follow-up and outcome evaluation

Patients were followed up with a face-to-face interview at 3 months. Clinical data were collected using an electronic data capture system by trained research coordinators based on a standardized interview protocol. The clinical outcome of this study was poor functional outcome, defined as an mRS ≥ 3 within 3 months confirmed by the treating hospital after AIS onset.

Model algorithms

We use ML models as follows to predict poor functional outcomes after 3 months in young patients:

- The XGBoost: is a scalable ML system for trees supported by Chen Tianqi of Washington College in 2016. The system runs more than ten times faster than existing popular solutions on a single machine and scales to millions of instances in distributed or memory-based environments [17].
- The GBTD: is a regression tree created using the gradient boosting method. It uses the gradient descent method and the function before loss is the squared error [20].
- The RF: is a combination of various trees identified by Leo Breiman et al. in 2001. It allows variables to be selected in the design, making it more robust to overcome the overfitting problem in the estimation [22].
- The LightGBM: It is an open-source library developed by Microsoft. It extends the gradient boosting algorithm with automatic selection and focuses on boosting samples with large gradients [18].
- Logistic regression: is a statistical method used to analyze the relationship between a categorical dependent variable and one or more independent variables. It predicts the probability of a categorical outcome, making it a powerful tool for understanding and predicting binary outcomes

Statistical analysis

Continuous variables were reported as mean and Standard deviation and group differences were tested for differences using the *t* test and the Mann–Whitney *U* test. In contrast, categorical variables are expressed as the number of cases and percentage divided by the number of cases, excluding missing and unknown cases. Fisher's exact test or the χ^2 test was used to compare categorical variables. Following the preliminary data, all patients were randomly divided into training and testing in a ratio of 80:20. The first group included 2070 patients with good functional outcomes (mRS ≤ 2), and the second group included 198 patients with poor functional outcomes (mRS ≥ 3). Tenfold cross-validation was used for feature selection and parameter fitting on the training set. The training process is used for modeling, while light testing is used only for model evaluation. We use the calibration plots to evaluate the calibration, the SFS technique for feature selection, and the GridSearch CV for hyperparameter tuning. The tuned hyperparameters for the ML models are listed in Supplementary Table 2.

The primary evaluation metric for discrimination was the AUC, while accuracy, positive predictive value (PPV), negative predictive value (NPV), and F1 score were considered secondary metrics. The differences between the logistic regression and other ML models were tested using the Delong test. The calibration of the models was evaluated using calibration curve plots. Statistical analysis was performed using SAS software (SAS 9.4) and Python software (python v3.9.7). Two-sided probability values < 0.05 are considered significant.

Results

Baseline characteristics

A total of 2268 young patients were included in our study after excluding 12,874 patients from 15,166, who are older than 50 years, and patients with missing data, as presented in Fig. 1. We had 24 variables with missing values, but the missing proportions for all variables are less than 5%, as shown in Supplemental Table 3. The mean age of our included patients was 44.3 ± 5.5 years, and 1787 (76%) patients were male. Table 1 presents the clinical characteristics of the young patients grouped into good and poor functional outcomes. After 3 months, 2070 patients (91%) had a good functional outcome (mRS 0–2) and 198 patients (9%) had poor functional outcomes (mRS 0–5), the mean age in each group was 44.3 ± 5.5 and 44.5 ± 5.3 respectively. The rate of males was 78.6% and 80.3% in the two groups respectively. Poor functional outcome in young patients was associated with numerous factors notably: living alone condition ($P = 0.06$), marital status ($P = 0.002$), education level ($P = 0.08$), history of

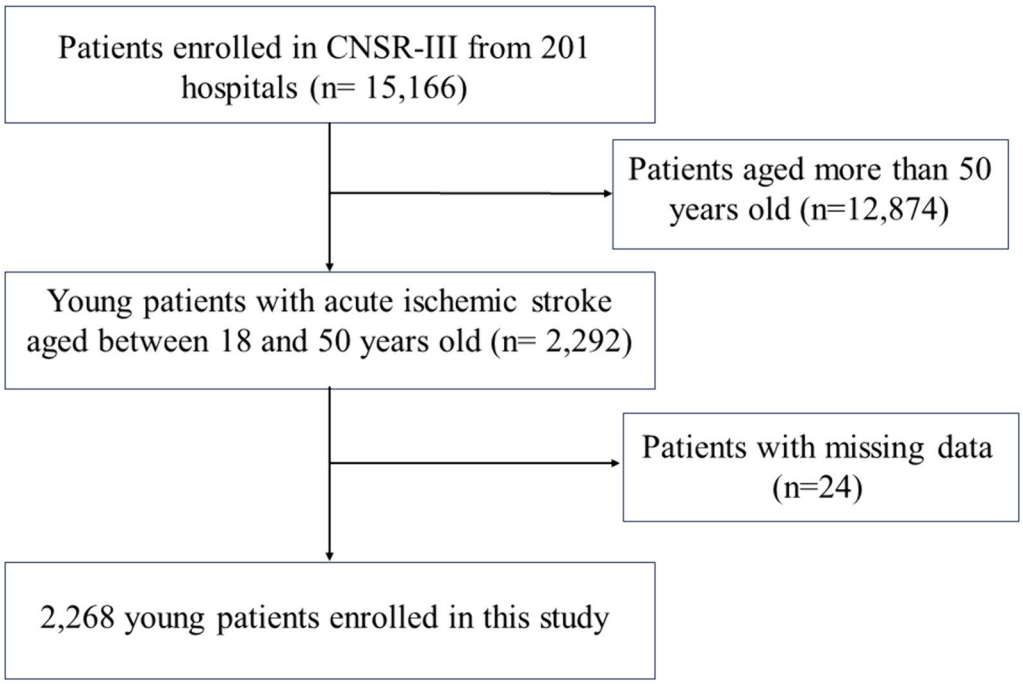

Fig. 1 Flow chart of the patients included in the study

smoking ($P=0.02$), heavy drinking ($P=0.06$) and stroke ($P<0.001$), including arterial atrial fibrillation ($P=0.02$), in addition to NIHSS score ($P<0.001$) and mRS score in admission ($P<0.001$) TOAST classification ($P<0.001$), secondary prevention, and laboratory values of lymphocytes ($P<0.001$) and neutrophils ($P<0.001$).

Feature selection

The feature selection was made by tenfold cross-validation and SFS–KNN. For the feature selection, choosing mRS in admission, living alone, and NIHSS in discharge remained independent predictors of a 3-month poor outcome (Supplementary Fig. 1).

Performance of the model

Table 2 shows the AUC, accuracy, PPV, NPV, and F1 scores of the different models. The XGBboost model achieved the highest AUC (AUC=0.801), followed by GBDT (AUC=0.795), RF (AUC=0.794), LightGBM (AUC=0.792), and logistic regression (AUC=0.789). The ROC curve and AUC of each machine-learning method compared with the logistic model are shown in Fig. 2. The predictive performance of the XGBoost ($P=0.03$), RF ($P=0.01$), and LightGBM ($P=0.04$) models was better than the logistic regression model. The calibration plots curve of XGBoost, GBDT, RF, and lightGBM are shown in Supplementary Fig. 2.

Discussion

In this study, the XGBoost model was identified as the optimal predictive model. This study aimed to determine the factors that lead to poor functional outcomes 3 months after an AIS in young patients. It also aimed to compare the predictive performance of the ML algorithm and the logistic model. The main findings of this study were:

- In 2268 young patients, poor functional outcome was significantly associated with a high mRS score at admission, living alone conditions, and a high NIHSS score at discharge.
- ML is superior to logistic regression, with XGBoost being the best model.

The lifelong impact of stroke in young adults is associated with significant costs for patients themself, their families, and society. The long-term medical, psychosocial, and socioeconomic consequences are particularly severe at younger ages [23]. Therefore, there is a need to identify risk factors and develop and validate predictive scores for post-AIS outcomes. Recently, many ML models have been designed to predict adverse outcomes using algorithms that can learn from large amounts of complex data. In a recent study, the RF method using a combination of Random Under-Sampling (RUS) and biomarkers was found to be the best stroke prediction

Table 1 Clinical characteristics of young patients according to modified Rankin score

Variables		All patients, 2268 (100%)	mRS (0–2),2070 (91%)	mRS (3–5), 198 (9%)	*P*
Age (years), mean ± SD		44.3 ± 5.5	44.3 ± 5.5	44.5 ± 5.3	0.68
Male, *n* (%)		1787 (79)	1628 (78.6)	159 (80.3)	0.59
BMI, kg/m_2		25.6 ± 3.6	25.6 ± 3.5	25.3 ± 3.7	0.17
Living Alone, *n* (%)		2185 (96)	1999 (96.6)	186 (93.9)	**0.06**
Marital, *n* (%)					**0.002**
Unmarried		53 (2.3)	47 (2.3)	6 (3.0)	
Married		2172 (95.8)	1990 (96.1)	182 (91.9)	
Divorced		35 (1.5)	27 (1.3)	8 (4.0)	
Widowed		4 (0.2)	2 (0.1)	2 (1.0)	
Remarried		1 (0.0)	1 (0.0)	0 (0.0)	
Education, *n* (%)					**0.08**
University		357 (15.7)	338 (16.3)	19 (9.6)	
High school		551 (24.3)	500 (24.2)	51 (25.8)	
Junior school		775 (34.2)	711 (34.3)	64 (32.3)	
Primary school		265 (11.7)	236 (11.4)	29 (14.6)	
Illiterate		51 (2.2)	44 (2.1)	7 (3.5)	
Alteplase, *n* (%)		169 (7.5)	148 (7.1)	21 (10.6)	0.08
Cigarette smoking, *n* (%)					**0.02**
Never		925 (40.8)	852 (41.2)	73 (36.9)	
Occasional		118 (5.2)	103 (5.0)	15 (7.6)	
Current		1031 (45.5)	948 (45.8)	83 (41.9)	
Former		194 (45.5)	167 (8.1)	27 (13.6)	
Alcohol drinking, *n* (%)					**0.06**
Never		976 (43.0)	900 (43.5)	76 (38.4)	
Occasional		686 (30.2)	630 (30.4)	56 (28.3)	
Current		464 (20.5)	418 (20.2)	46 (23.2)	
Former		142 (6.3)	122 (5.9)	20 (10.1)	
mRS in admission, *n* (%)					**< 0.001**
0		1842 (81.2)	367 (17.7)	13 (6.6)	
1		290 (12.8)	811 (39.2)	23 (11.6)	
2		69 (3.0)	389 (18.8)	22 (11.1)	
3		28 (1.2)	241 (11.6)	33 (16.7)	
4		32 (1.4)	250 (12.1)	90 (45.5)	
5		7 (0.3)	12 (0.6)	17 (8.6)	
Initial NIHSS, mean ± SD		2.1 ± 3.0	3.3 ± 3.3	8.4 ± 5.6	**< 0.001**
Medical History, *n* (%)					
Hypertension		1222 (53.9)	1103 (53.3)	119 (60.1)	**0.07**
Diabetes		377 (16.6)	343 (16.6)	34 (17.2)	0.83
Dyslipidemia		191 (8.4)	176 (8.5)	15 (7.6)	0.65
Migraine		43 (1.9)	42 (2.0)	1 (0.5)	0.26
Stroke		271 (11.9)	231 (11.2)	40 (20.2)	**< 0.001**
TIA		68 (3.0)	65 (3.1)	3 (1.5)	0.20
Heart diseases		122 (5.4)	110 (5.3)	12 (6.1)	0.66
Arterial fibrillation		35 (1.5)	28 (1.4)	7 (3.5)	**0.02**
Family history, *n* (%)					
Hypertension	681 (30.0)		626 (30.2)	55 (27.8)	0.48
Diabetes	212 (9.3)		190 (9.2)	22 (11.1)	0.32
Dyslipidemia	69 (3.0)		65 (3.1)	4 (2.0)	0.66
Stroke	418 (18.4)		375 (18.1)	43 (21.7)	0.46

Table 1 (continued)

Variables		All patients, 2268 (100%)	mRS (0–2),2070 (91%)	mRS (3–5), 198 (9%)	*P*
Heart Diseases	418 (18.4)		134 (6.5)	10 (5.1)	0.71
Cancer	81(3.6)		77 (3.7)	4 (2.0)	0.29
TOASTClassification, *n* (%)					**< 0.001**
LAA		506 (22.3)	438 (21.2)	68 (34.3)	
Cardiac embolism		51 (2.2)	43 (2.1)	8 (4.0)	
Small-vessel occlusion		494 (21.8)	472 (22.8)	22 (11.1)	
Other determined cause		50 (2.2)	45 (2.2)	5 (2.5)	
Undetermined		1167 (51.5)	1072 (51.8)	95 (48.0)	
Laboratory tests, mean ± SD					
Lymphocyte (10^9/L)		2.0 ± 0.7	2.0 ± 0.7	1.8 ± 0.7	**< 0.001**
Neutrophile (10^9/L)		5.3 ± 2.8	5.2 ± 2.6	6.6 ± 3.5	**< 0.001**
Platelet (10^9/L)		238.6 ± 65.5	238.7 ± 65.1	237.6 ± 69.8	0.81
Cholesterol (mmol/l)		4.3 ± 1.3	4.3 ± 1.3	4.3 ± 1.3	0.51
HDL (mmol/l)		1.1 ± 0.3	1.1 ± 0.3	1.1 ± 0.3	0.64
LDL (mmol/l)		2.5 ± 1.0	2.5 ± 1.0	2.6 ± 1.1	0.20
Triglyceride (mmol/l)		1.9 ± 1.4	1.9 ± 1.5	1.7 ± 1.0	**0.06**
Creatinine (μmol/L)		72.5 ± 33.8	72.5 ± 34.3	72.8 ± 28.6	0.92
Uric acid (μmol/L)		325.9 ± 93.9	326.8 ± 93.0	316.1 ± 102.6	0.14
Treatment at discharge, n (%)					
Antiplatelet therapy		2079 (91.7)	1910 (92.3)	169 (85.4)	**< 0.001**
Anticoagulant therapy		40 (1.8)	34 (1.6)	6 (3.0)	**< 0.001**
Statin therapy		2047 (90.3)	1879 (90.8)	168 (84.8)	**< 0.001**
NIHSS at Discharge, mean ± SD		2.1 ± 3.0	1.7 ± 2.1	7.2 ± 5.5	**< 0.001**

Alteplase: intravenous thrombolysis with alteplase; BMI: Body mass index; HDL: High-density lipoprotein; LAA: large-artery atherosclerosis; LDL: Low-density lipoprotein; mRS: modified Rankin Scale; NIHSS: National Institutes of Health Stroke Scale; *P*: significant value (> 0.1); TIA: transit ischemic attack; TOAST: The Trial of Org 10,172 in Acute Stroke Treatment; Cigarette Smoking: Never: Never smoked; Occasional: Smoke, but have not smoked at least 1 cigarette per day for more than 1 year; Current: average of at least 1 cigarette per day for 1 year before the onset of disease; Former: Smoke at least 1 cigarette per day for more than 1 year in a lifetime. However, quit smoking 1 year before stroke onset. Alcohol drinking: Never: Never drink alcohol; Occasional: Drinks alcohol but has not had at least 1 drink per week for more than 1 year; Current: Drinking alcohol at least once a week for more than 1 year before the onset of the disease; Former: At least 1 drink per week for more than 1 year in a lifetime. However, abstained from alcohol for 1 year before stroke onset.

Table 2 Test sets result of machine learning models at 3-month stroke outcome prediction

Model	AUC	Accuracy	PPV	NPV	F1-score
XGboost	**0.801**	**0.923**	**0.591**	**0.939**	**0.412**
GBDT	0.795	0.923	0.6	0.937	0.406
Random Forest	0.794	0.909	0.468	0.943	0.422
LightGBM	0.792	0.918	0.541	0.937	0.412
Logistic regression	0.789	0.923	0.667	0.930	0.314

XGboost: eXtreme Gradient Boosting; GBDT: Gradient Boosting Decision Trees; LightGBM: Light Gradient Boosted Machine Algorithm, AUC: Area Under Curve; PPV: positive predictive value; NPV: negative predictive value

model in Chinese adult patients with hypertension [24]. A multidisciplinary study of atherosclerosis found that of nine predictive tests, RF was the best model for predicting cardiovascular disease risk including AIS [25].

In addition, results from the China Longitudinal Health and Longevity Study, show that red light running (RLR) applied to the Synthetic Minority Over-sampling Technique (SMOTE) is superior to other test models in predicting stroke in the elderly [22]. Also, the study by Hao et al. showed that a deep neural network model could improve the prediction of long-term outcomes in 2604 AIS patien ts aged 66.2 ± 12.6 years [9].

Our study shows that the XGBoost model has good discrimination (AUC = 0.81), and is better than other algorithms in predicting poor functional outcomes in young AIS patients within 3 months, followed by RF, lightGBM, and GBDT. Among them, XGBoost, RF, and lightGBM were better than logistic regression. Choosing the right ML model for disease prediction is critical for optimization. Various ML models have already been developed to predict clinical outcomes after stroke in both general and elderly patients. The study of Chen

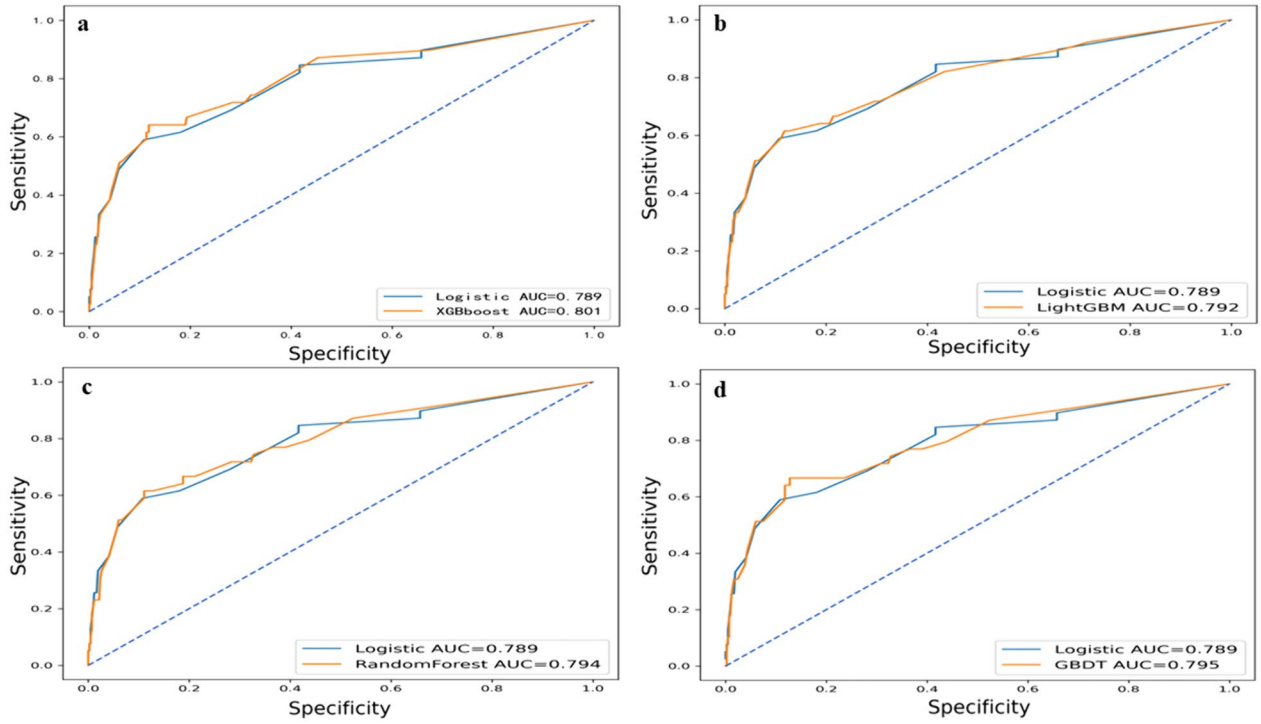

Fig. 2 Receiver Operating Characteristic curves for machine learning models. a: ROC curve of XGBoost model; b: ROC curve of LightGBM model; c: ROC curve of Random Forest model; d: ROC curve of GBDT model

et al. suggested that the CatBoost algorithm had the best predictive performance compared to logistic regression and other ML models [14], and found that gender, age, stroke history, heart rate, D-dimer, creatinine, TOAST classification, mRS at admission and discharge, and NIHSS score at discharge predicts poor outcomes at 90 days in patients with TIA [14]. In addition, the study by Xiang et al. [26] showed that the RF model could better predict 6-month outcomes of Chinese AIS patients than the Houston intra-arterial therapy (HIAT) score, the total health risks in vascular events (THRIVE) score, as well, the NIHSS score on admission, age, previous Diabetes mellitus and crEatinine (NADE) Nomogram. This study found that NIHSS at admission, age, premorbid mRS, fasting glucose, and creatinine were significant predictor factors. Moreover, the study by Xio et al. proved that the XGB model is a reliable predictive model, and also showed that hypertension, cancer, congestive heart failure, chronic lung, and peripheral vascular disease may be closely associated with stroke in elderly patients [27]. However, predicting risk factors for poor functional impairment in young patients using different types of ML remains unclear.

Feature selection from ML has shown that a high mRS score at admission and a high NIHSS score at discharge, as well as, the patient living alone remained independent predictors of poor 3-month outcomes in young patients with AIS. The NIHSS and mRS are quantitative tools used to efficiently and effectively assess the degree of neurological impairment in patients with AIS. In addition, these neurological severity scores are closely related to the patient's brain necrosis volume, location, type, perfusion, and injury [28, 29]. On the other hand, our results are consistent with Waje-Andreassen et al. who found that living alone was a predictor of long-term mortality in 232 young stroke patients [30]. Additionally, in the Riks-Stroke-based study, living alone condition was an independent predictor of short-term mortality after stroke [31]. In addition, a recent study suggests that stroke severity is associated with living alone [32]. Mathew et al. showed that individuals living alone at home were much less likely to arrive at the hospital early than those living with others and that this delay resulted in a much lower thrombolysis treatment rate [33]. The Swedish stroke registry also showed that treatment rates were ≈ 50% lower in patients living alone [34], which may explain the association between the condition of living alone and poor functional outcomes after stroke in young patients. Moreover, other studies have demonstrated that living alone can be considered a proxy for low social support, and for coronary heart

disease, biological processes such as inflammatory and prothrombotic disorders, and mental disorders [35].

Our finding shows that the XGboost model can better predict the risk of 3-month poor functional outcomes in young patients with AIS. These results are similar to the study by Chung et al. which suggested that the XGBoost model is a reliable predictive power for AIS and also demonstrates the validity of the model for use in patients receiving various AIS treatments [15]. In addition, Yuan et al. have shown that the XGBoost model has better performance in predicting the 90-day readmission risk in AIS patients [36]. XGBoost is a new integrated learning method that boosts gradient. It implements a ML algorithm in the context of gradient boosting and is efficient, flexible, and portable. XGBoost is an efficient gradient-boosting algorithm capable of handling large-scale datasets, outperforming many other ML algorithms in terms of performance. It features built-in regularization, effectively preventing overfitting and enhancing the model's generalization ability. Overall, XGBoost excels in processing large-scale data, high-dimensional features, and complex tasks. The XGBoost classification method is more suitable for clinical predictive analysis than other ML techniques because it is effective and can combine the classification and regression tree process, allowing the processing of different, complex, and nonlinear models (such as multiple cases, and medical conditions). The potential of ML to significantly improve health care by automating routine processes and improving clinical decision-making is tantalizing today [37]. The future is likely to be characterized by augmented intelligence, in which computers become indispensable tools for patient care, and allow physicians to spend more time on patient care [38].

In the future, we could use the XGBoost model accessible via an online web page or integrated into clinical decision support systems (CDSS). This would allow clinicians to conveniently use the model in their daily work. Additionally, providing clear expectations to patients and their families can help them better understand the illness and actively participate in the treatment and rehabilitation process. Our prediction model will require further validation in prospective studies to confirm its effectiveness. We believe that with additional research and validation, the XGBoost model has the potential to be widely applied in clinical practice, enhancing the treatment and prognosis of young stroke patients.

Using the smallest variables to achieve better predictions is our strength. The simpler a model is, the easier it is to validate. Second, the predictors used in our study were comprehensive and included demographic, lifestyle, and clinical variables, which allowed us to examine the relationship between risk factors and stroke from multiple perspectives. In addition, the data used in this study were from large a Chinese cohort with high-quality data representing AIS patients in China.

Our study also has some limitations, first, there is some level of missing values, but, our all-missing values are < 5%. We used imputations to fill in missing laboratory data, and no statistical differences were observed between the data before and after the imputation process. Second, this study does not include genomic and imaging data, which may have limited predictive power. Third, external validation is absent, and this will be conducted in an independent external cohort population in the future.

Conclusion

Our results suggest that employing ML methods particularly XGBoost may improve upon conventional logistic regression models in identifying young stroke patients at risk of poor functional outcomes within 3 months.

Abbreviations

AIS	Acute Ischemic Stroke
AUC	Area Under The Receiving Curve
DNN	Deep Neural Network
XGBoost	Extreme Gradient Boosting
HIAT	Houston intra-arterial therapy
ML	Machine learning
mRS	Modified Rankin Scale
NPV	Negative Predictive Value
NADE	NIHSS score on admission, age, previous Diabetes mellitus and crEatinine
PPV	Positive Predictive Value
RF	Random Forest
RUS	Random Under-Sampling
RLR	Red Light Running
SFS	Sequential Forward Selection
SVM	Support Vector Machine
SMOT	Synthetic Minority Over-Sampling Technique
GBTD	The Gradient Boosting Decision Trees
KNN	The K Nearest Neighbors
LightGBM	The Light Gradient Boosted Machine
TOAST	The Trial Of ORG 10172 In Acute Stroke Treatment
NIHSS	The National Institutes Of Health Stroke Scale
CNSR-III	Third Chinese National Stroke Registry
TIA	Transient Ischemic Attack

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-024-02056-3.

Supplementary materials 1: Figure S1. Feature selection plots. The *x*-axis is labeled "Number of features" suggesting that individual features of a dataset are being added incrementally. The *y*-axis is labeled "Performance" with values ranging from approximately 0.72 to 0.84.

Supplementary materials 2: Figure S2. Calibration plots for prediction of stroke outcome at 3 months across test sets The *x*-axis is labeled " Predicted probability "; The *y*-axis is labeled " Fraction of positives". a: Calibration plot of XGBoost model; b: Calibration plot of LightGBM model; c: Calibration plot of Random Forest model; d: Calibration plot of GBDT model.

Supplementary materials 3.

Supplementary materials 4.

Supplementary materials 5.

Acknowledgements
We express our gratitude to the Changping Laboratory for their invaluable support. We extend our sincere appreciation to all the participating hospitals, doctors and nurses, and the members of the Third China National Stroke Registry Steering Committee members, particularly Dr Yongjun Wang and Dr Yong Jiang, for their unwavering support and assistance.

Author contributions
Lamia Mbarek Writing–Original Draft Preparation, Conceptualization; Siding Chen Data Curation, Data Analysis, Methodology,; Aoming Jin and Yuesong Pan Visualization Validation; Zhe Xu Validation; Xia Meng and Xiaomeng Yang Project Administration; investigation; Yong Jiang Project Administration; Writing–Review & Editing; Supervision; Yongjun Wang Writing–Review & Editing, Supervision, Funding Acquisition, Investigation.

Funding
This study was supported by grants from the National Natural Science Foundation of China (U20A20358), the Capital's Funds for Health Improvement and Research (2020–1-2041), and the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2019-I2M-5–029).

Data availability
The data supporting the results of this study are available upon reasonable request from the corresponding author. No datasets were generated or analysed during the current study.

Declarations

Ethical approval and consent to participate
The CNSR-III was approved by the Ethics Committee at Beijing Tiantan Hospital (IRB approval number: KY2015-001–01) and all participating centers. It was conducted following the Declaration of Helsinki (2013 revision). All participants had informed consent from the patient or legally authorized representative (primarily spouse, parents, adult children, otherwise indicated).

Consent for publications
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China. [2]China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, No.119 South 4th Ring West Road, Fengtai District, Beijing 100070, China. [3]Changping Laboratory, Beijing, China. [4]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University and Capital Medical University, Beijing 100091, China. [5]Research Unit of Artificial Intelligence in Cerebrovascular Disease, Chinese Academy of Medical Sciences, Beijing 2019RU018, China. [6]Beijing Advanced Innovation Centre for Big Data-Based Precision Medicine, Beihang University, Capital Medical University, Beijing, China. [7]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China. [8]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China.

Received: 28 May 2024 Accepted: 9 September 2024
Published online: 10 October 2024

References

1. Wang YJ, Li ZX, Gu HQ, Zhai Y, Zhou Q, Jiang Y, et al. China stroke statistics: an update on the 2019 report from the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. Stroke Vasc Neurol. 2022;7(5):415–50.
2. Wang YJ, Li ZX, Gu HQ, Zhai Y, Jiang Y, Zhao XQ, et al. China Stroke Statistics 2019: a Report From the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. Stroke Vasc Neurol. 2020;5(3):211–39.
3. Ekker MS, Boot EM, Singhal AB, Tan KS, Debette S, Tuladhar AM, et al. Epidemiology, aetiology, and management of ischaemic stroke in young adults. Lancet Neurol. 2018;17(9):790–801.
4. George MG. Risk factors for ischemic stroke in younger adults: a focused update. Stroke. 2020;51(3):729–35.
5. Ekker MS, Verhoeven JI, Schellekens MMI, Boot EM, Van Alebeek ME, Brouwers PJAM, et al. Risk factors and causes of ischemic stroke in 1322 young adults. Stroke. 2023;54(2):439–47.
6. Feigin VL, Roth GA, Naghavi M, Parmar P, Krishnamurthi R, Chugh S, et al. Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet Neurol. 2016;15(9):913–24.
7. Xian Y, Thomas L, Liang L, Federspiel JJ, Webb LE, Bushnell CD, et al. Unexplained variation for hospitals' use of inpatient rehabilitation and skilled nursing facilities after an acute ischemic stroke. Stroke. 2017;48(10):2836–42.
8. Putaala J. Ischemic stroke in the young: current perspectives on incidence, risk factors, and cardiovascular prognosis. Eur Stroke J. 2016;1(1):28–40.
9. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. Stroke. 2019;50(5):1263–5.
10. Saber H, Somai M, Rajah GB, Scalzo F, Liebeskind DS. Predictive analytics and machine learning in stroke and neurovascular medicine. Neurol Res. 2019;41(8):681–90.
11. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188:kwz189.
12. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, et al. Random forest-based prediction of stroke outcome. Sci Rep. 2021;11(1):10071.
13. Forkert ND, Verleger T, Cheng B, Thomalla G, Hilgetag CC, Fiehler J. Multiclass support vector machine-based lesion mapping predicts functional outcome in ischemic stroke patients. Baron JC, editor. PLoS ONE. 2015;10(6):e0129569.
14. Chen SD, You J, Yang XM, Gu HQ, Huang XY, Liu H, et al. Machine learning is an effective method to predict the 90-day prognosis of patients with transient ischemic attack and minor stroke. BMC Med Res Methodol. 2022;22(1):195.
15. Chung CC, Su ECY, Chen JH, Chen YT, Kuo CY. XGBoost-based simple three-item model accurately predicts outcomes of acute Ischemic stroke. Diagnostics. 2023;13(5):842.
16. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022;6(12):1330–45.
17. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.Accessed 17 Oct 2023.
18. Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol. 2021;22(1):271.
19. Reis I, Baron D, Shahaf S. Probabilistic random forest: a machine learning algorithm for noisy data sets. Astron J. 2019;157(1):16.
20. Peng T, Chen X, Wan M, Jin L, Wang X, Du X, et al. The prediction of hepatitis E through ensemble learning. Int J Environ Res Public Health. 2020;18(1):159.

21. Wang Y, Jing J, Meng X, Pan Y, Wang Y, Zhao 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(3):158–64.
22. Wu Y, Fang Y. Stroke prediction with machine learning methods among older Chinese. Int J Environ Res Public Health. 2020;17(6):1828.
23. Maaijwee NAMM, Rutten-Jacobs LCA, Arntz RM, Schaapsmeerders P, Schoonderwaldt HC, Van Dijk EJ, et al. Long-term increased risk of unemployment after young stroke: a long-term follow-up study. Neurology. 2014;83(13):1132–8.
24. Huang X, Cao T, Chen L, Li J, Tan Z, Xu B, et al. Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults. Front Cardiovasc Med. 2022;6(9): 901240.
25. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multiethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
26. Li X, Pan X, Jiang C, Wu M, Liu Y, Wang F, et al. Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning. Front Neurol. 2020;19(11): 539509.
27. Zhang X, Fei N, Zhang X, Wang Q, Fang Z. Machine learning prediction models for postoperative stroke in elderly patients: analyses of the MIMIC database. Front Aging Neurosci. 2022;18(14): 897611.
28. Zöllner JP, Misselwitz B, Kaps M, Stein M, Konczalla J, Roth C, et al. National Institutes of Health Stroke Scale (NIHSS) on admission predicts acute symptomatic seizure risk in ischemic stroke: a population-based study involving 135,117 cases. Sci Rep. 2020;10(1):3779.
29. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. Stroke. 2007;38(3):1091–6.
30. Waje-Andreassen U, Naess H, Thomassen L, Eide GE, Vedeler CA. Long-term mortality among young ischemic stroke patients in western Norway. Acta Neurol Scand. 2007;116(3):150–6.
31. Lindmark A, Glader EL, Asplund K, Norrving B, Eriksson M. For the Riks-StrokeCollaboration. Socioeconomic disparities in stroke case fatality–observations from riks-stroke, the Swedish stroke register. Int J Stroke. 2014;9(4):429–36.
32. Aron AW, Staff I, Fortunato G, McCullough LD. Prestroke living situation and depression contribute to initial stroke severity and stroke recovery. J Stroke Cerebrovasc Dis. 2015;24(2):492–9.
33. Reeves MJ, Prager M, Fang J, Stamplecoski M, Kapral MK. Impact of living alone on the care and outcomes of patients with acute stroke. Stroke. 2014;45(10):3083–5.
34. Eriksson M, Jonsson F, Appelros P, Åsberg KH, Norrving B, Stegmayr B, et al. Dissemination of thrombolysis for acute ischemic stroke across a nation: experiences from the swedish stroke register, 2003 to 2008. Stroke. 2010;41(6):1115–22.
35. Reblin M, Uchino BN. Social and emotional support and its implication for health. Curr Opin Psychiatry. 2008;21(2):201–5.
36. Xu Y, Yang X, Huang H, Peng C, Ge Y, Wu H, et al. Extreme gradient boosting model has a better performance in predicting the risk of 90-Day readmissions in patients with ischaemic stroke. J Stroke Cerebrovasc Dis. 2019;28(12): 104441.
37. Caballé-Cervigón N, Castillo-Sequera JL, Gómez-Pulido JA, Gómez-Pulido JM, Polo-Luque ML. Machine learning applied to diagnosis of human diseases: a systematic review. Appl Sci. 2020;10(15):5135.
38. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. Can Med Assoc J. 2021;193(34):E1351–7.

Publisher's Note