BMC Medical Research Methodology

RESEARCH Open Access

# Predictive etiological classification of acute ischemic stroke through interpretable machine learning algorithms: a multicenter, prospective cohort study

Siding Chen[1,2,3], Xiaomeng Yang[1], Hongqiu Gu[1,2], Yanzhao Wang[4], Zhe Xu[1,2], Yong Jiang[1,2,3,5*] and Yongjun Wang[1,2,3,6,7,8*]

## Abstract

**Background** The prognosis, recurrence rates, and secondary prevention strategies varied significantly among different subtypes of acute ischemic stroke (AIS). Machine learning (ML) techniques can uncover intricate, non-linear relationships within medical data, enabling the identification of factors associated with etiological classification. However, there is currently a lack of research utilizing ML algorithms for predicting AIS etiology.

**Objective** We aimed to use interpretable ML algorithms to develop AIS etiology prediction models, identify critical factors in etiology classification, and enhance existing clinical categorization.

**Methods** This study involved patients with the Third China National Stroke Registry (CNSR-III). Nine models, which included Natural Gradient Boosting (NGBoost), Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (LGBM), Gradient Boosting Decision Tree (GBDT), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), and logistic regression (LR), were employed to predict large artery atherosclerosis (LAA), small vessel occlusion (SVO), and cardioembolism (CE) using an 80:20 randomly split training and test set. We designed an SFS-XGB with 10-fold cross-validation for feature selection. The primary evaluation metrics for the models included the area under the receiver operating characteristic curve (AUC) for discrimination and the Brier score (or calibration plots) for calibration.

**Results** A total of 5,213 patients were included, comprising 2,471 (47.4%) with LAA, 2,153 (41.3%) with SVO, and 589 (11.3%) with CE. In both LAA and SVO models, the AUC values of the ML models were significantly higher than that of the LR model ($P < 0.001$). The optimal model for predicting SVO (AUC [RF model] = 0.932) outperformed the optimal LAA model (AUC [NGB model] = 0.917) and the optimal CE model (AUC [LGBM model] = 0.846). Each model displayed relatively satisfactory calibration. Further analysis showed that the optimal CE model could identify potential CE patients in the undetermined etiology (SUE) group, accounting for 1,900 out of 4,156 (45.7%).

*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

**Conclusions** The ML algorithm effectively classified patients with LAA, SVO, and CE, demonstrating superior classification performance compared to the LR model. The optimal ML model can identify potential CE patients among SUE patients. These newly identified predictive factors may complement the existing etiological classification system, enabling clinicians to promptly categorize stroke patients' etiology and initiate optimal strategies for secondary prevention.

**Keywords** Acute ischemic stroke, Clinical prediction, Etiological classification, Prospective cohort study, Machine learning

## Introduction

Stroke is the second leading cause of global mortality and the primary contributor to both morbidity and disability in China. Acute ischemic stroke (AIS) represents a prevalent form of stroke [1–3]. Different subtypes of AIS have varying prognostic trajectories, recurrence patterns, and strategies for secondary prevention. Accurate identification of AIS subtypes is pivotal for developing effective secondary prevention strategies and alleviating the burden associated with AIS.

The most widely accepted AIS subtyping system is the Trial of ORG 10,172 in Acute Stroke Treatment (TOAST) classification scheme [4]. However, the initial assessment of AIS is often time-consuming and uncertain, requiring expert reviewers to thoroughly interpret clinical indicators, conduct laboratory tests, and analyze electrocardiography and imaging results [5, 6]. This process is highly dependent on the expertise and experience of the doctors [7, 8]. Despite rigorous training, physicians frequently encounter challenges in identifying AIS subtypes. Reports also indicate that primary psychiatrists exhibit low accuracy in assessing AIS etiology due to the requirement for extensive experience accumulation [5, 6, 9]. Therefore, it is crucial to develop rapid etiological classification prediction models that can accurately identify AIS subtypes during the acute stage after admission.

Recent advancements in machine learning (ML) applications across various healthcare domains have sparked innovations in developing novel ML-based etiological classification technologies [10, 11]. The non-parametric nature of ML and its ability to capture non-linear relationships make it well-suited for identifying AIS subtypes, given the complex and non-normal nature of most medical data. Studies have demonstrated the potential of ML in this field. For instance, one study used ML algorithms to automatically identify and quantify carotid artery plaques in MRI scans, achieving 91.41% accuracy in LAA classification using Random Forest (RF) [12]. Wang et al. developed a predictive model for patients with large vessel occlusion using an RF model, achieving an area under the receiver operating characteristic curve (AUC) of 0.831 [13]. This study concluded that ML outperformed logistic regression (LR) in identifying patients with large vessel occlusion. Sun et al. employed various ML algorithms to develop an etiological prediction model for large artery atherosclerosis (LAA) using 62 features [14]. However, these studies have limitations such as small sample sizes, single-center retrospective designs, and poor interpretability.

To address these limitations, it is imperative to use large prospective cohort data and advanced ML algorithms to develop more accurate etiological prediction models with fewer predictive variables. This study aims to develop predictive models for LAA, small vessel occlusion (SVO), and cardiogenic embolism (CE) using interpretable ML algorithms based on high-quality, prospective cohort studies to provide explanations for predictive factors and complement existing etiological classifications.

## Methods

### Study design and participants

We extracted data from the Third China National Stroke Registry (CNSR-III), a large-scale nationwide prospective registry of acute ischemic cerebrovascular events in China. The study design and patient identification details for CNSR-III have been reported previously [2]. Imaging data were collected in the Digital Imaging and Communications in Medicine (DICOM) format on discs and interpreted by trained professional physicians. Stroke subtypes were classified into five major categories according to TOAST classification: LAA, SVO, CE, other determined etiology (SOE), and undetermined etiology (SUE) [4]. Additionally, our data included the Causative Classification System (CCS), which integrates etiological and phenotypic classifications [15].

A total of 44 biomarkers identified in this study were extracted from these samples. We excluded 4,190 patients without baseline plasma, serum, or imaging data and also excluded patients with SOE from the analysis. Ultimately, we enrolled 5,213 patients (including LAA, SVO, and CE) for the main study and included 4,156 SUE patients for subsequent analysis (Fig. 1).

### Data information

Based on published literature and pathophysiological considerations, the candidate variables included in our study comprised demographic characteristics, medical history, family history, and imaging and laboratory data. Detailed information can be found in the supplementary materials (Table S1).

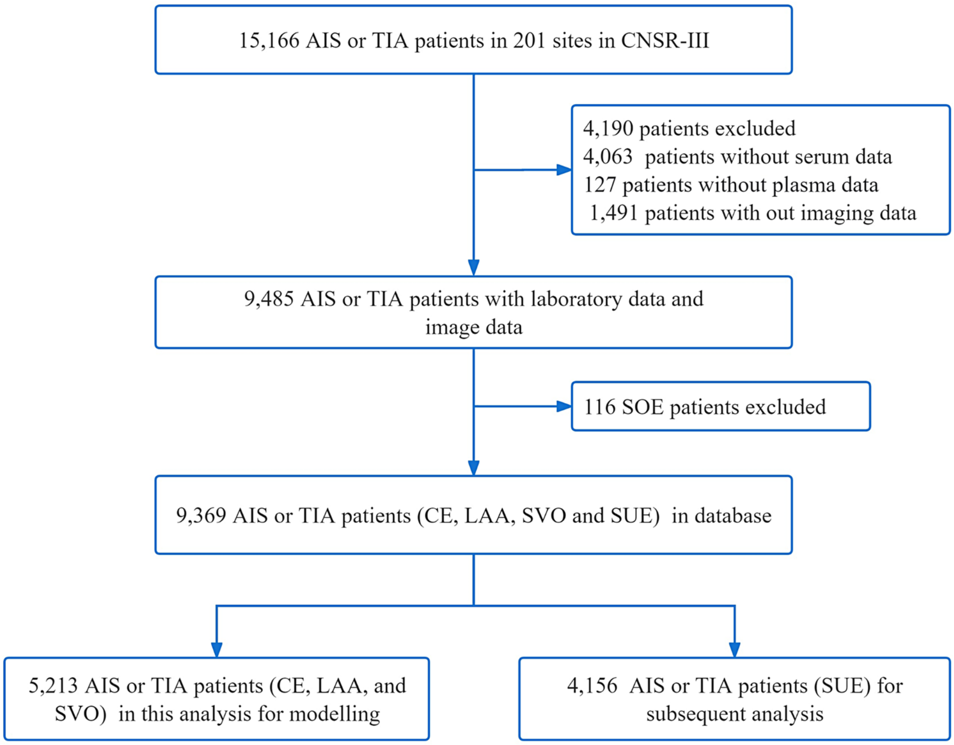
15,166 AIS or TIA patients in 201 sites in CNSR-III
4,190 patients excluded
4,063 patients without serum data
127 patients without plasma data
1,491 patients with out imaging data
9,485 AIS or TIA patients with laboratory data and image data
116 SOE patients excluded
9,369 AIS or TIA patients (CE, LAA, SVO and SUE) in database
5,213 AIS or TIA patients (CE, LAA, and SVO) in this analysis for modelling
4,156 AIS or TIA patients (SUE) for subsequent analysis

**Fig. 1** Study flowchart

## ML algorithms

In this study, we developed and compared eight ML predictive models to assess their performance against the LR model. The models included Natural Gradient Boosting (NGBoost [NGB]) [16], Light Gradient Boosting Machine (LGBM) [17], Categorical Boosting (CatBoost [CAT]) [18], Extreme Gradient Boosting (XGBoost [XGB]) [19], Gradient Boosting Decision Tree (GBDT) [20, 21], Random Forest (RF) [22, 23], Adaptive Boosting (AdaBoost [Ada]) [24, 25], and Support Vector Machine (SVM) [26]. The dataset was randomly divided into a training set (80%) and a testing set (20%). We used 10-fold cross-validation for parameter optimization in the training set. Details of the parameters for different algorithms can be found in the supplementary materials (Table S6).

(1) NGB: NGB is a novel algorithm for regression prediction tasks. It extends conventional gradient boosting algorithms by incorporating natural gradients to optimize model parameters, enhancing its ability to adapt to the probability distribution characteristics of the data. The primary objective of NGB is to directly model the predictive distribution, moving beyond mere predictions of expected values [16].

(2) LGBM: LGBM is a robust gradient-boosting framework known for its computational efficiency. Compared to traditional gradient-boosting decision tree algorithms, LGBM offers faster training speeds and lower memory consumption [17].

(3) CAT: CAT is an innovative ordered gradient boosting algorithm that utilizes ordered target-based statistics to handle categorical features and employs permutation strategies to prevent prediction shifts [18].

(4) XGB: XGB is a robust ML algorithm used for classification and regression problems. It enhances gradient-boosting trees by combining multiple decision trees to improve predictive capabilities [19].

(5) GBDT: GBDT employs the gradient descent method to reduce error. This model can automatically capture interactions between features without the need for manually specifying interaction terms and is relatively robust to outliers and noisy data [20].

(6) RF: RF is an ensemble supervised learning method consisting of multiple decision trees, each trained on different subsets of the data. The results from each tree are averaged, which reduces variance and improves predictive performance [22].

(7) Ada: Ada is an iterative ensemble learning method. Its core idea is to combine multiple weak learners, typically weak classifiers like decision trees or Naive Bayes, to create a strong learner [24].

(8) SVM: SVM is a robust algorithm used for classification tasks. It finds the optimal hyperplane that maximizes the margin between classes,

ensuring effective separation of data points. SVM is particularly effective in high-dimensional spaces and can handle non-linear classification using various kernel functions [26].

### Features selections

We employed our custom-designed Sequential Forward Selection with XGB (SFS-XGB), utilizing 10-fold cross-validation to maximize performance. Within the training set, we implemented 10-fold cross-validation with SFS, varying the parameter k from 3 to 10. The optimal feature set was evaluated based on AUC values. From the SFS-XGB results, we identified the top 10 variables as candidates. Our objective was to pinpoint the optimal feature set with the highest AUC values while minimizing the number of variables. This approach was applied separately to identify the best predictive feature subsets for LAA, SVO, and CE. Notably, to ensure specificity in CE models—especially concerning conditions like atrial fibrillation (AF)—we excluded medical histories of AF and heart valve disease (HVD) during feature selection for CE models.

### Data preprocessing

For data preprocessing, we employed multiple imputations to complete missing values in continuous variables from laboratory data, while categorical variables were imputed using the mode. The distribution of laboratory data was evaluated both before and after imputation, with detailed statistics provided in Table S2.

To standardize laboratory data, we utilized MinMaxScaler, which linearly transformed the features, scaling them to fit within the [0, 1] range.

Addressing the imbalance in our dataset, particularly for the classification of CE against other categories (LAA and SVO), where the sample ratio was 589:4,624 = 1:7.893, indicating significant class imbalance, we applied random undersampling to the training set using the imblearn library. The sampling strategy was set to achieve a balanced ratio, specifically 0.5.

### Subsequent analysis

In our subsequent analysis, we divided it into two parts: 1) We utilized our best models for LAA, SVO, and CE to identify potential patients within the SUE group. Using an 80% probability cutoff, patients below this threshold were categorized as UND, while those above were selected. The group with the highest probability was considered the final prediction group.2) We extended our analysis by applying the LAA, SVO, and CE models to further study the CCS subtypes. Based on their AUC performance, we assigned weights ranging from 0 to 1 to the top four models in each etiology prediction category. ML scores were calculated by multiplying these weights by a total score of 12, as detailed in Table S10. In cases where a model was duplicated, scores were summed. By combining predictors from LAA, SVO, and CE, we created a comprehensive set of predictors.

### Definitions of metrics

We evaluated our models' performance using discrimination and calibration as primary measures. Discrimination, measured by the AUC, indicates the model's ability to distinguish, with higher values indicating better performance. Calibration was assessed using the Brier score, which ranges from 0 to 1, with a lower score indicating better calibration [27]. Calibration plots were also utilized for visual assessment.

Additional metrics used in this study included accuracy, sensitivity, specificity, Youden's index, and F1-score. These metrics utilize True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to describe correct and incorrect predictions of unknown etiology types. The calculations for these measures were as follows:

$$\mathrm{accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad (1)$$

$$\mathrm{sensitivity} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (2)$$

$$\mathrm{specificity} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (3)$$

$$\mathrm{Youden\prime\,s\ index} = \mathrm{sensitivity} + \mathrm{specificity} - 1 \quad (4)$$

$$\mathrm{F1\ score} = 2\,(\mathrm{precision}^{*}\mathrm{sensitivity})\,/(\mathrm{precision} + \mathrm{sensitivity}) \quad (5)$$

### Statistical analysis

Baseline characteristics were presented using means and standard deviations or medians and interquartile ranges for continuous variables, and frequencies and percentages for categorical variables. The chi-square test or Fisher's exact test was used to compare baseline characteristics among categorical variables, while analysis of variance (ANOVA) or the Kruskal-Wallis test was employed for continuous variables. Differences in AUC values among various models were assessed using the DeLong test [28], and model interpretations were facilitated using SHapley Additive exPlanations (SHAP) [29]. The suitability of ML research was evaluated based on TRIPOD and PROBAST guidelines [30, 31]. Data analysis was performed using SAS software (version 9.4) and Python (version 3.9.7). All comparisons were two-sided, with statistical significance defined as $P < 0.05$.

## Results

### Baseline characteristics

From an initial cohort of 15,166 patients with AIS or transient ischaemic attack (TIA), 4,190 patients without serum, plasma, or imaging data were excluded, leaving 9,485 patients for analysis (Fig. 1). Among these, 5,213 patients diagnosed with LAA, SVO, and CE were included. The distribution among these groups was as follows: 47.4% (*n*=2,471) were LAA, 41.3% (*n*=2,153) were SVO, and 11.3% (*n*=589) were CE. The average age was 62.9±11.1 years, with 30.0% (*n*=1,563) females. The median (IQR) admission NIHSS score was 3.0 (2.0–6.0). The most prevalent medical history was hypertension (64.8%, *n*=3,380), followed by diabetes (24.4%, *n*=1,273) and prior stroke (23.4%, *n*=1,218). More than half of the patients presented with a single infarction (51.7%, *n*=2,697) or anterior circulation infarction (56.6%, *n*=2,951). Demographic details for the LAA, SVO, and CE groups can be found in Table S1.

### Feature selection of LAA, SVO, and CE models

The dataset was randomly divided into training and testing sets at an 80:20 ratio. The training set initially comprised 70 variables from Table S1. As shown in Table S2, there was no statistically significant difference between the data before and after multiple imputation of missing values ($P>0.1$).

Feature selection was exclusively conducted in the training set, and the process is detailed in Fig. 2 and Tables S3-S5. For the LAA models, seven features were selected: number of acute infarctions, history of AF, blood glucose level, age, longitude, admission NIHSS score, and total cholesterol (CHOL). The SVO models utilized ten variables: age, number of acute infarctions, history of AF, infarction circulation, admission NIHSS score, C-reactive protein (hs-CRP), absolute lymphocyte count (LYM), low-density lipoprotein (LDL-C), smoking history, and history of diabetes. In CE models, AF and the history of HVD demonstrated strong discriminatory power (Figure S1). Optimal performance for CE models was achieved with six features: history of heart disease

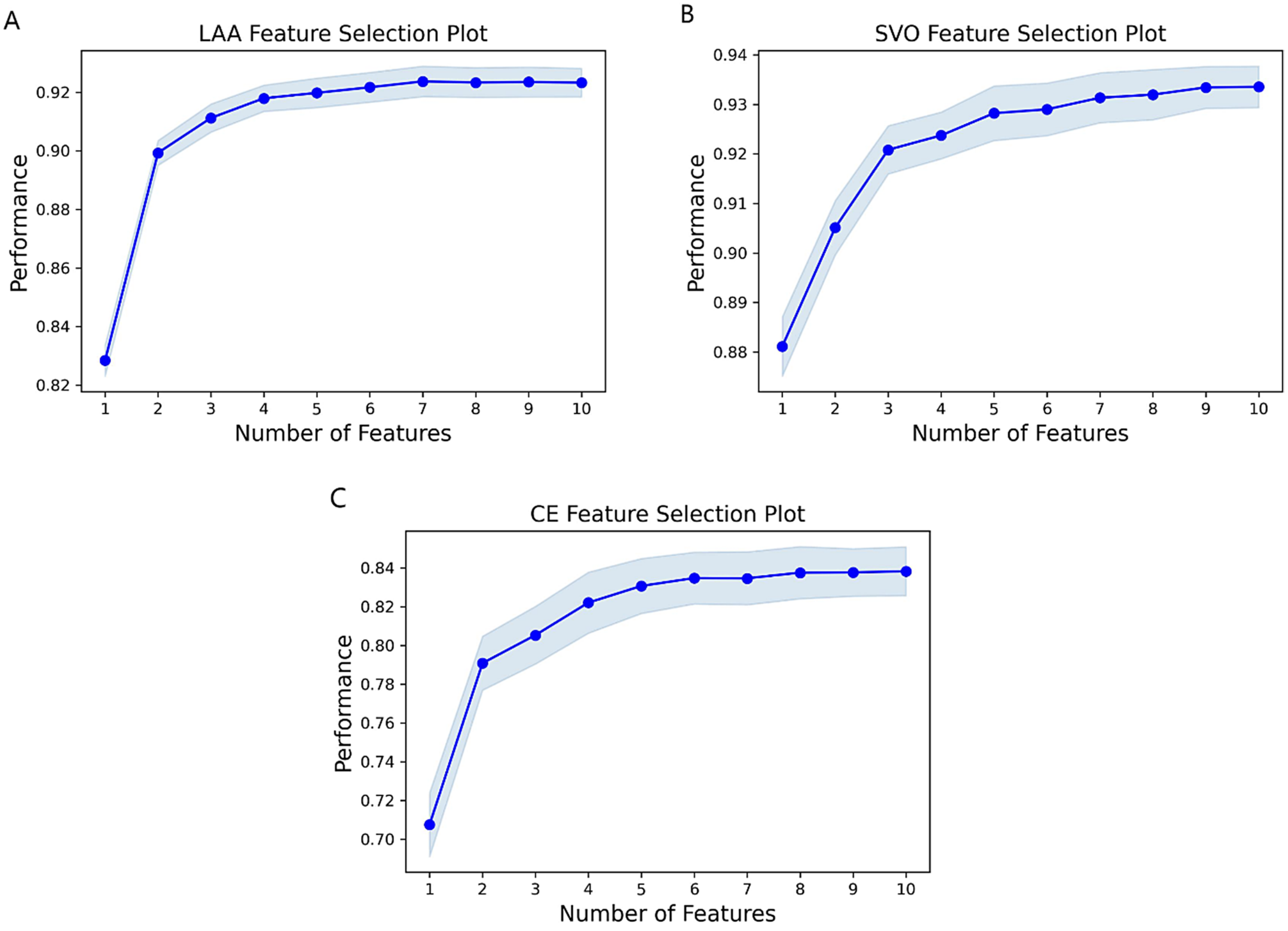

**Fig. 2** Feature selection plots of LAA, SVO and CE models (in the training set). A, the Feature selection diagram of the LAA etiology classification prediction models; B, the Feature selection diagram of the SVO etiology classification prediction models; C, the Feature selection diagram of the CE etiology classification prediction models

**Table 1** Comparative analysis results of nine etiology prediction models (in the test set)

| Models | AUC | Accuracy | Sensitivity | Specificity | F1 score | Youden's index | Brier score |
|---|---|---|---|---|---|---|---|
| **LAA** | | | | | | | |
| NGB | 0.917 | 0.881 | 0.789 | 0.964 | 0.863 | 0.753 | 0.096 |
| RF | 0.916 | 0.881 | 0.789 | 0.964 | 0.863 | 0.753 | 0.112 |
| CAT | 0.913 | 0.881 | 0.789 | 0.964 | 0.863 | 0.753 | 0.097 |
| LGBM | 0.913 | 0.881 | 0.789 | 0.964 | 0.863 | 0.753 | 0.097 |
| XGB | 0.912 | 0.881 | 0.789 | 0.964 | 0.863 | 0.753 | 0.098 |
| Ada | 0.912 | 0.880 | 0.796 | 0.956 | 0.863 | 0.752 | 0.170 |
| GBDT | 0.910 | 0.878 | 0.789 | 0.958 | 0.860 | 0.748 | 0.099 |
| SVM | 0.897 | 0.882 | 0.796 | 0.960 | 0.865 | 0.755 | 0.098 |
| LR | 0.841 | 0.849 | 0.713 | 0.973 | 0.818 | 0.68 | 0.151 |
| **SVO** | | | | | | | |
| RF | 0.932 | 0.892 | 0.988 | 0.824 | 0.883 | 0.812 | 0.110 |
| LGBM | 0.930 | 0.894 | 0.995 | 0.822 | 0.885 | 0.817 | 0.083 |
| GBDT | 0.930 | 0.882 | 0.956 | 0.830 | 0.870 | 0.786 | 0.086 |
| NGB | 0.929 | 0.893 | 0.995 | 0.820 | 0.885 | 0.816 | 0.084 |
| XGB | 0.929 | 0.890 | 0.984 | 0.824 | 0.881 | 0.807 | 0.085 |
| CAT | 0.928 | 0.895 | 0.993 | 0.825 | 0.886 | 0.818 | 0.084 |
| Ada | 0.924 | 0.882 | 0.956 | 0.830 | 0.870 | 0.786 | 0.151 |
| SVM | 0.900 | 0.891 | 0.981 | 0.827 | 0.881 | 0.808 | 0.091 |
| LR | 0.835 | 0.775 | 0.763 | 0.783 | 0.737 | 0.546 | 0.152 |
| **CE** | | | | | | | |
| LGBM | 0.846 | 0.895 | 0.576 | 0.935 | 0.553 | 0.511 | 0.091 |
| XGB | 0.846 | 0.890 | 0.585 | 0.929 | 0.545 | 0.513 | 0.093 |
| GBDT | 0.845 | 0.887 | 0.568 | 0.928 | 0.532 | 0.495 | 0.094 |
| CAT | 0.844 | 0.899 | 0.559 | 0.943 | 0.557 | 0.502 | 0.092 |
| NGB | 0.842 | 0.895 | 0.559 | 0.937 | 0.545 | 0.497 | 0.089 |
| LR | 0.842 | 0.904 | 0.559 | 0.948 | 0.569 | 0.507 | 0.090 |
| RF | 0.839 | 0.897 | 0.559 | 0.941 | 0.552 | 0.500 | 0.094 |
| Ada | 0.832 | 0.877 | 0.585 | 0.915 | 0.519 | 0.499 | 0.181 |
| SVM | 0.780 | 0.893 | 0.492 | 0.944 | 0.509 | 0.435 | 0.103 |

Abbreviations: AUC, area under the curve; LR, logistic regression; RF, random forests; XGB, extreme gradient boosting; NGB, Natural Gradient Boosting; LGBM, Light Gradient Boosting Machine; CAT, Categorical Boosting; GBDT, Gradient Boosting Decision Tree; Ada, Adaptive Boosting; SVM, Support Vector Machine

AUC [LAA-CAT]=0.91324, AUC [LAA-LGBM]=0.91308, AUC [LAA-XGB]=0.91205, AUC [LAA-Ada]=0.91196; AUC [SVO-LGBM]=0.92992, AUC [SVO-GBDT]=0.92953, AUC [SVO-NGB]=0.92903, AUC [SVO-XGB]=0.92864; AUC [CE-LGBM]=0.84601, AUC [CE-XGB]=0.84553, AUC [CE-NGB]=0.84162, AUC [CE-LR]=0.84150

(HD), age, history of coronary heart disease (CHD), direct bilirubin (DBIL), adiponectin, and international normalized ratio (INR).

### Model construction and evaluation

In the training set, we optimized parameters for constructing nine prediction models, detailed in Table S6. Table 1 was presented the performance metrics for each model evaluated on the test set. Among the nine models, the ML models outperformed the LR model. Specifically, the NGB model excelled in predicting LAA, while the RF model performed best in predicting SVO. Additionally, the LGBM model demonstrated superior efficacy in predicting CE. The AUC values of ML models for LAA and SVO predictions were significantly better than those of the LR model according to Delong's test results with a p-value less than 0.001. For CE predictions, although not reaching statistical significance based on Delong's test results with a p-value greater than 0.05, the AUC performance of LGBM, XGB, GBDT, CAT, and NGB models surpassed the LR model. The ROC curves presented in Fig. 3 illustrate the performance of prediction models for LAA, SVO, and CE, respectively. Among them, SVO models exhibited superior performance followed by LAA models and then CE models. All these predictive models demonstrated excellent calibration, as evidenced by the calibration curves shown in Fig. 4 and Figures S8-S10.

### Visualization of feature importance

SHAP was employed to illustrate our LAA, SVO, and CE models (Fig. 5). This plot visualized the relationship between feature values and SHAP values in the test set, highlighting higher SHAP values as indicators of greater influence on classification under each etiology. Dependence plots (Figure S2-S4) further elucidated the impact

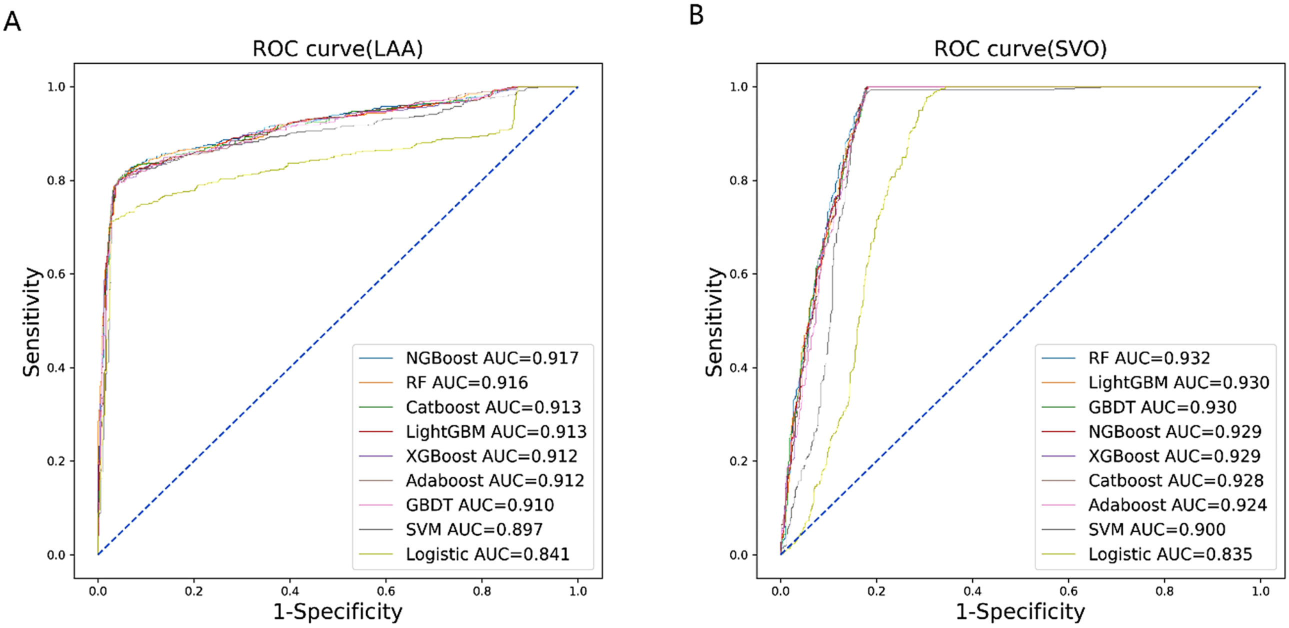

**Fig. 3** ROC curves plots of LAA, SVO and CE models (in the test set). A, the ROC curves of the LAA etiology classification prediction models; B, the ROC curves of the SVO etiology classification prediction models; C, the ROC curves of the CE etiology classification prediction models

of the single feature on the output of the etiological classification models.

As shown in Fig. 5, the contributions of each predictor variable to LAA, SVO, and CE models were highlighted. The number of infarctions and history of AF were identified as the most significant variables in LAA and SVO models. Admission NIHSS score and age were found to be common predictors for both models. Optimal variable combinations for classifying LAA, SVO, and CE patients were illustrated in partial dependence plots (Figure S2-S4).

## Subsequent analysis

(1) Additional analysis for SUE: A total of 4156 SUE patients were included (Fig. 1), with a mean age of 61.9 ± 11.4 years and 1369 females (32.9%). The distribution of different genders in SUE was presented in Table S8. The established optimal

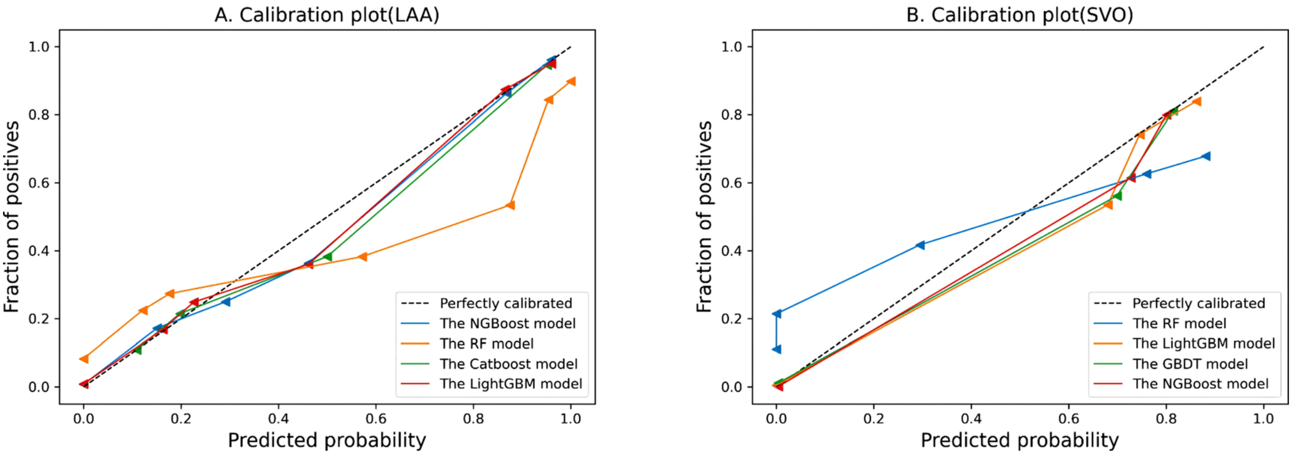

**Fig. 4** Calibration curves for the top four models in each etiology classification model (in the test set). A, the calibration curves of the LAA etiology classification prediction models; B, the calibration curves of the SVO etiology classification prediction models

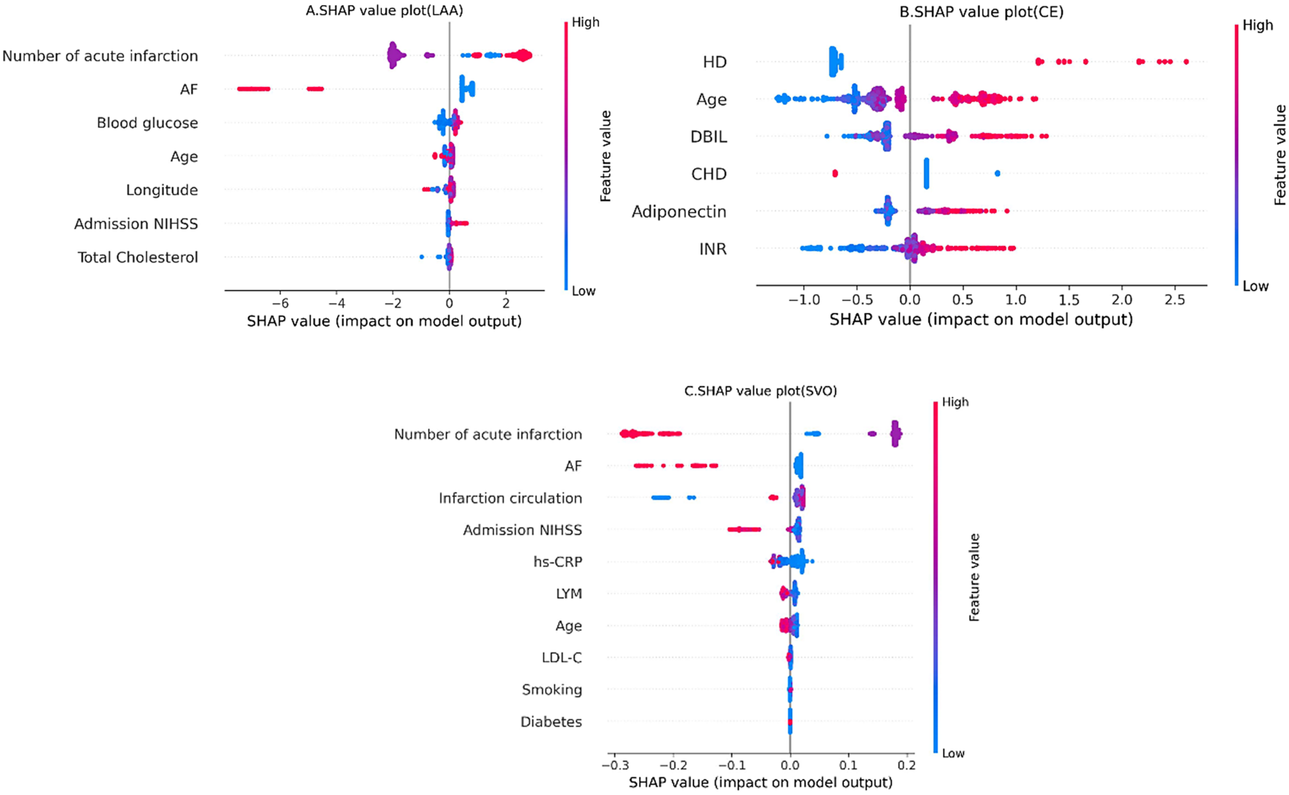

**Fig. 5** SHAP summary plot of the LAA, SVO and CE models (in the test set). A, the SHAP summary plot of the LAA etiology classification prediction models; B, the SHAP summary plot of the CE etiology classification prediction models; C, the SHAP summary plot of the SVO etiology classification prediction models

LAA (NGB), SVO (RF), and CE (LGBM) models were utilized to identify potential LAA, SVO, or CE patients in SUE. Results indicated that 1900 (45.7%) potential CE patients could be identified in SUE, with 2256 SUE patients (UND) having a predicted probability below 80%. Detailed results can be found in Table S7 and Figure S5. We compared 1900 potential CE and UND within SUE and observed statistically significant differences ($P<0.05$) in several heart-related variables between these two groups (Table S9).

(2) Extended analysis for CCS: Out of 4,642 patients in the CCS classification, there were 2,163 LAA patients, 2,020 SVO patients, and 459 CE patients. In the test set, there were 433 LAA patients, 404 SVO patients, and 92 CE patients. For CCS analysis,

predictors such as age, smoking, admission NIHSS score, longitude, diabetes, AF, HD, CHD, infarction circulation, number of acute infarctions, blood glucose, INR, CHOL, LDL-C, hs-CRP, DBIL, Adiponectin, and LYM were used. Models were scored as follows: RF (18), LGBM (16.8), NGB (15.6), GBDT (10.2), CAT (4.92), and XGB (1.2) (Table S10). Subsequent analysis combined the top 2 models (RF, LGBM) with these 18 variables. In the test set, the RF model accurately predicted 392 LAA (90.5%), 72 CE (78.3%), and 404 SVO (100.0%). LGBM correctly predicted 393 LAA (90.8%), 72 CE (78.3%), and 404 SVO (100.0%). The highest accuracy in predicting SVO was demonstrated by ML models, followed by LAA and CE (Figure S6-S7).

## Discussion

To our knowledge, this study was the first application of ML algorithms for classifying AIS etiology within a prospective high-quality Chinese AIS cohort. Notably, this study also marks the initial utilization of the NGB algorithm in this specific field. We comprehensively integrated clinical, imaging, and laboratory data to accurately classify AIS subtypes. The ML algorithms successfully constructed predictive models for LAA, SVO, and CE, demonstrating robustness consistent with findings in other fields [10, 11, 13, 32, 33]. Among the developed models, the SVO model showed superior performance, followed by the LAA and CE models. It is worth noting that the top-performing models for etiological classification were LAA-NGB, SVO-RF, and CE-LGBM. Our CE-LGBM model successfully identified 1,900 (45.7%) potential CE patients in SUE. Our study revealed the following clinical findings:

(1) Individuals aged 57–70 with multiple infarcts, high NIHSS scores (> 8), no history of AF, elevated blood glucose (> 6 mmol/L), and high CHOL levels (≥ 5 mmol/L) in regions like Northeast China, North China, and East China were more likely to develop LAA than CE or SVO.
(2) Those without a history of AF, under 61 years old, with low NIHSS scores (< 7), a single infarct in one circulation (anterior or posterior), and smokers (maintaining low lymphocyte, hs-CRP, and LDL-C levels) were more likely to develop SVO instead of LAA or CE.
(3) Besides strong CE indicators like AF and HVD, older age (> 69 years), history of HD, impaired coagulation (INR > 1.15), no CHD history, and elevated DBIL (> 5μmol/L) and adiponectin (> 2.5 mg/ml) levels indicated a higher likelihood of developing CE rather than LAA or SVO.

LAA significantly contributes to disability and mortality in China [34], commonly associated with risk factors like high cholesterol, hypertension, smoking, diabetes, and older [35]. SVO results from small blood vessel occlusion or narrowing, leading to limited blood supply to the brain. It typically presents mild symptoms like dizziness and limb numbness. Compared to LAA and CE, SVO has a better prognosis due to its limited impact on smaller brain areas [36]. CE is characterized by heart-related clots that travel to the brain, causing cerebral vascular embolism and subsequent ischemic injury. It is associated with conditions such as AF, rheumatic heart disease, and heart valve issues. CE has a less favorable prognosis and a higher disability rate compared to other stroke types. Accurate identification of CE is crucial for personalized treatment, often involving anticoagulant medications. This study attempted to identify additional factors to distinguish CE from other subtypes, utilizing interpretable ML models to aid clinical decision-making. Therefore, this is one of the reasons why this study separately constructed three etiological prediction models of LAA, CE, and SVO.

Age, NIHSS score, AF, and CHD history were common factors influencing stroke classification [4]. It's essential to note that a fasting blood glucose level below 5.6 mmol/L was considered normal, but our study suggests that LAA risk increases with blood glucose levels exceeding 6 mmol/L. Elevated blood glucose levels may contribute to LAA stroke risk through several mechanisms: (1) The promotion of atherosclerosis occurs through the damage to arterial endothelial cells, inflammation induction, and encouragement of cholesterol and lipid accumulation in arterial walls. This leads to plaque formation and narrowing of arteries; (2) hyperglycemia increases the risk of platelet aggregation and coagulation, promoting thrombus formation and embolism in narrowed arteries, contributing to LAA; (3) high blood glucose levels affect arterial wall elasticity, inflammation, and oxidative stress, potentially damaging arterial endothelial cells, accelerating atherosclerosis, and increasing LAA risk [37–39]. The NIHSS score provides valuable insights into the clinical symptoms and neurological conditions of LAA and SVO, but it does not specifically indicate the etiological subtype. Therefore, combining the NIHSS score with other clinical and imaging data is crucial for a comprehensive evaluation of LAA. Individuals born in Northeast China, North China, and East China showed higher susceptibility to LAA compared to other subtypes. Various factors, such as dietary habits, natural environmental factors, and genetic influences, may have contributed to this heightened susceptibility. For instance, the prevalence of high-fat and high-salt diets in Northeast China could have elevated the risk of hypertension and hyperlipidemia, potentially leading to arterial wall damage and lipid deposition, thereby increasing the likelihood of atherosclerosis [40–43]. However, further research and validation are needed to fully understand the specific impact of different regions on stroke etiology classification.

To our knowledge, our study was the first to utilize ML in discovering adiponectin and DBIL as a potential novel biomarker associated with CE. Adiponectin was a peptide substance secreted from adipose tissue with anti-inflammatory and anti-atherosclerotic effects [44]. Previous studies have linked it to various diseases, including energy metabolism [45], immune response, chronic inflammatory conditions [46], and atherosclerosis [47, 48]. Low adiponectin levels may make individuals more susceptible to LAA rather than CE. Additionally, elevated adiponectin levels could serve as a biomarker for CE, indicating underlying biological mechanisms that warrant further investigation. Elevated DBIL levels are indicative of liver and biliary system disorders. High DBIL levels have been associated with increased stroke severity and poorer prognosis [49]. However, the role of DBIL as a stroke risk factor or prognostic indicator remained uncertain due to potential confounding factors [50]. Further research is needed to establish the causal relationship between DBIL and CE. Elevated INR indicated prolonged coagulation time in patients, likely due to the frequent use of anticoagulant medications in individuals with CE, thus displaying this distinctive characteristic.

Our feature selection utilized SFS-XGB with 10-fold cross-validation, optimizing predictor selection based on AUC performance. This method effectively removed irrelevant features, reduced dimensionality, and enhanced model accuracy. Addressing the imbalance between CE and other stroke subtypes (LAA and SVO) through random undersampling ensured reliable model training, mitigating bias towards the majority class and improving prediction reliability. The existing reports and guidelines provide support for AF and HVD as high-risk cardiogenic sources of CE [4, 51–53], which aligns with our results (Table S1 and Fig. 1). To uncover other variables linked to CE etiology classification, a strategy was adopted that excluded atrial fibrillation and heart valve disease.

Among our LAA prediction models, NGB demonstrated the highest performance (AUC=0.917), closely followed by RF (AUC=0.916). Importantly, all ML models significantly outperformed the LR model (*P*<0.001). Notably, the NGB model exhibited superior calibration compared to RF (Brier score: NGB=0.096, RF=0.112), further validating its predictive accuracy. As far as we know, this study also represents the first attempt to investigate the use of the NGB model for predicting stroke etiological classification. NGB was developed by the ML team at Stanford in 2019; NGB is a boosting algorithm designed to provide probabilistic forecasts through a full probability distribution rather than point predictions [16]. Here's a detailed breakdown of the NGB algorithm, including the mathematical formalisms: For many ML models, we seek to optimize the parameters $\theta$ to minimise loss $L\left(y, f\left(x;\theta\right)\right)$ where $y$ was the target and $f\left(x;\theta\right)$ was the prediction. The gradient boosting algorithm improved the model iteratively by fitting new base models to the negative gradient of the loss with respect to the current prediction. Instead of point predictions, consider a distribution $P_\theta\left(y|x\right)$ parameterized by $\theta$ to represent the prediction. The objective was to minimise the expected value of some scoring rule $S\left(y, P_\theta\left(y|x\right)\right)$. A common choice for $S$ was the negative log-likelihood: $S\left(y, P_\theta\left(y|x\right)\right) = -\log P_\theta\left(y|x\right)$. NGB generalises gradient boosting to parameterized probability distributions [16]. The updated to $\theta$ were done using the natural gradient instead of the gradient. Given the scoring rule $S$, the steepest descent direction (natural gradient) was given by:

$$g\left(\theta\right) = I^{-1}\left(\theta\right)\nabla_\theta S\left(y, P_\theta\left(y|x\right)\right)$$

where $I\left(\theta\right)$ is the Fisher Information matrix.

NGB overcame the challenge of probabilistic predictions with gradient boosting, showing high accuracy in predicting structured or tabular data and excelling in LAA etiology prediction. Despite its advantages, NGB did not guarantee superior performance in all scenarios, with the RF model's classification effect being only slightly lower (difference of 0.001). The RF model performed best in predicting SVO (AUC=0.932), closely followed by LGBM (AUC=0.930). All ML models significantly outperformed the LR model (*P*<0.001). RF's robustness in LAA and SVO models stemmed from constructing multiple decision trees based on Gini Impurity or Information Gain, random sampling to prevent overfitting, and displaying good noise resistance and fast training speed.

LGBM showed the best predictive performance for CE models (AUC=0.846). Known for its speed and suitability for large-scale datasets, LGBM employs efficient strategies like leaf-wise tree growth and histogram-based training. Future improvements could involve integrating heart-related examination variables to enhance CE differentiation. Although our study did not fully utilize LGBM's potential due to feature selection constraints, it holds promise for improved performance in larger datasets. Additionally, excluding correlated factors like AF and HVD in the initial CE model might have affected overall performance compared to SVO and LAA models. Diagnosing CE is complex, requiring thorough cardiogenic source identification and high-risk factor consideration.

Subsequent research could explore LGBM further for enhanced results.

Our analysis identified 45.7% (1900) of SUE patients as potential CE patients (Table S7), showing significant differences in heart disease-related variables compared to the remaining UND group ($P<0.05$, Table S9). Precise ML models effectively identified LAA, SVO, and CE patients among SUE, easing diagnostic challenges and improving treatment accuracy. We ranked RF, LGBM, NGB, GBDT, CAT, and XGB as top performers. RF and LGBM, particularly in SVO predictions, demonstrated high accuracy within the CCS system. This aligns with our initial models' performance, indicating their robustness. Regarding the selection of these three algorithms, we would like to offer the following recommendations:

(1) Each algorithm came with its own set of hyperparameters. Proper tuning was crucial for optimal performance. An improperly tuned NGB might underperform compared to a well-tuned LGBM or RF.
(2) RF combined predictions from multiple decision trees to produce a more robust and accurate outcome.
(3) NGB focused on probabilistic predictions and used natural gradients. If the problem did not require probabilistic forecasting, the added complexity might not be beneficial.
(4) LGBM was efficient and scalable, especially for large datasets. When dealing with substantial data, using LGBM is advisable.
(5) The choice of algorithm should be based on the nature of the data, the specific problem context, and thorough experimentation and validation.

Despite data normalization efforts, the SVM model underperformed in predicting LAA, SVO, and CE compared to other ML models. SVM's preference for limited samples and numerous features hindered its effectiveness with our larger sample size. Additionally, SVM's difficulty in finding suitable kernel functions and its focus on boundary data points resulted in lower AUC performance, highlighting its limitations relative to other models.

Our ML models demonstrated exceptional predictive efficiency while maintaining precision. Given the simplicity and accessibility of the variables used in this study's prediction model, these three etiological prediction models can be easily integrated into web pages or clinical decision support systems (CDSS) for practical application. This will enable clinicians to efficiently classify patients' etiological factors. By utilizing ML algorithms to identify new variables, we have filled gaps in existing clinical knowledge regarding variable selection. We firmly believe that there is no universally superior method; the key lies in selecting the appropriate algorithm and variables for specific clinical challenges. Although using three etiological prediction models may seem more complex than using a single model in clinical practice, it is important to note that core predictors differ among patients with different etiological subtypes. Therefore, segregating the prediction of these subtypes could lead to improved accuracy for each subtype. Additionally, our three prediction models can also be employed to forecast potential LAA, SVO, and CE patients within the SUE population. The ability of our model to identify potential CE patients among those with SUE has significant implications in clinical practice as it addresses the challenge of delayed anticoagulation treatment due to ambiguous etiological diagnoses. In our subsequent analysis section, we found that the results of our LAA, SVO, and CE etiological prediction models were in good agreement with actual classifications. Future studies seeking to establish a singular predictive model can reference our discovered predictor variables when building their models. Both the TOAST classification and our models provide valuable insights for clinicians, facilitating precise patient assessment. By integrating our models with guidelines and clinical expertise, clinicians can thoroughly evaluate patients and implement optimal preventive or intervention measures that ultimately improve patient prognosis.

However, this study had several limitations. Firstly, the predictors used to establish CE models had limited ability to identify CE patients (AUC≤0.846). Future research should incorporate additional brain imaging, ECG, and echocardiographic data to uncover more relevant variables. Secondly, while ML algorithms, especially RF, LGBM, and NGB, showed high accuracy and AUC performance in SVO and LAA models, further external validation is essential. Suitable external validation data were not available in our existing databases due to the origin of predictor variables from clinical, imaging, and laboratory data. We plan to establish an appropriate cohort for external validation. Currently, we recommend applying the etiological prediction model to retrospective data while prospective prediction needs to be evaluated. Thirdly, this analysis used multiple imputations to handle missing values. The majority of missing variables accounted for less than 5%, but we couldn't confidently assert that variables with missing values exceeding 5% were randomly missing, as there was no direct method to test this [54]. Nonetheless, we minimized selection bias, and the data distribution after multiple imputations did not significantly differ from the distribution before imputations.

## Conclusions

In conclusion, our interpretable ML models, which combine clinical, imaging, and lab data, successfully classify patients with LAA, SVO, and CE, outperforming traditional LR models. Additionally, our model can identify potential CE patients within the SUE group, supplementing existing classifications. This potentially enables clinicians to promptly categorize stroke patients based on their etiologies and initiate optimal prevention and treatment strategies.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02331-1.

Supplementary Material 1

### 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. Additionally, we would like to acknowledge the diligent efforts of the editors and reviewers who provided meticulous feedback and constructive comments during the review process.

### Author contributions

SDC wrote the main text of the manuscript, performed the data analysis, and prepared the supplementary materials. YJ and YJW were responsible for supervision and data provision. XMY reviewed the article and provided clinical consultation. YZW contributed to the discussion of machine learning algorithms and offered theoretical assistance on machine learning algorithms. XMY, ZX, and HQG participated in the clinical discussion. All authors read and approved the manuscript for submission.

### Funding

This study was supported by grants from National Natural Science Foundation of China (U20A20358), the Capital's Funds for Health Improvement and Research (2020-1-2041) and Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2019-I2M-5-029).The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Data availability

No datasets were generated or analysed during the current study.

## Declarations

### Ethics approval and consent to participate

This study was approved by the Ethics Committees of Beijing Tiantan Hospital (IRB number: KY2015-001-01). Written informed consent was obtained from all participants or their representatives.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

[1]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, No.119 South 4th Ring West Road, Fengtai District, Beijing 100070, China
[2]China National Clinical Research Center for Neurological Diseases, No.119 South 4th Ring West Road, Fengtai District, Beijing 100070, China
[3]Changping Laboratory, Beijing, China
[4]School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China
[5]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University & Capital Medical University, Beijing 100091, China
[6]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China
[7]Clinical Center for Precision Medicine in Stroke, Capital Medical University, Beijing, China
[8]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

Received: 24 November 2023 / Accepted: 5 September 2024
Published online: 10 September 2024

### References

1. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. 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.
2. 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;4:158–64.
3. Wang Y-J, Li Z-X, Gu H-Q, Zhai Y, Jiang Y, Zhao X-Q, the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association. Stroke Vasc Neurol. 2020;5:211–39. 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 CollaborationsChina Stroke Statistics 2019: A Report From.
4. Adams HP, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24:35–41.
5. Yang X-L, Zhu D-S, Lv H-H, Huang X-X, Han Y, Wu S, et al. Etiological classification of cerebral ischemic stroke by the TOAST, SSS-TOAST, and ASCOD systems: the impact of Observer's experience on reliability. Neurologist. 2019;24:111–4.
6. Goldstein LB, Jones MR, Matchar DB, Edwards LJ, Hoff J, Chilukuri V, et al. Improving the reliability of stroke subgroup classification using the trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria. Stroke. 2001;32:1091–7.
7. Jauch EC, Barreto AD, Broderick JP, Char DM, Cucchiara BL, Devlin TG, et al. Biomarkers of Acute Stroke etiology (BASE) study methodology. Transl Stroke Res. 2017;8:424–8.
8. Hankey GJ. Secondary stroke prevention. Lancet Neurol. 2014;13:178–94.
9. Pandian JD, Kalkonde Y, Sebastian IA, Felix C, Urimubenshi G, Bosch J. Stroke systems of care in low-income and middle-income countries: challenges and opportunities. Lancet. 2020;396:1443–51.
10. 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;50:1263–5.
11. Kamel H, Navi BB, Parikh NS, Merkler AE, Okin PM, Devereux RB, et al. Machine learning prediction of stroke mechanism in Embolic strokes of undetermined source. Stroke. 2020;51:e203–10.
12. Latha S, Muthu P, Lai KW, Khalil A, Dhanalakshmi S. Performance Analysis of Machine Learning and Deep Learning architectures on early stroke detection using carotid artery ultrasound images. Front Aging Neurosci. 2022;13:828214.
13. 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:e001096.
14. Sun T-H, Wang C-C, Wu Y-L, Hsu K-C, Lee T-H. Machine learning approaches for biomarker discovery to predict large-artery atherosclerosis. Sci Rep. 2023;13:15139.
15. Ay H, Benner T, Murat Arsava E, Furie KL, Singhal AB, Jensen MB, et al. A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of Stroke System. Stroke. 2007;38:2979–84.

16. Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A et al. Ngboost: Natural gradient boosting for probabilistic prediction. In: International conference on machine learning. PMLR; 2020. pp. 2690–700.
17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30.
18. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. 2018.
19. 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. 2016. pp. 785–94.
20. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;:1189–232.
21. Peng T, Chen X, Wan M, Jin L, Wang X, Du X, et al. The prediction of Hepatitis E through Ensemble Learning. IJERPH. 2020;18:159.
22. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
23. Wang C, Chen X, Du L, Zhan Q, Yang T, Fang Z. Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease. Comput Methods Programs Biomed. 2020;188:105267.
24. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. Stat its Interface. 2009;2:349–60.
25. Tran BX, Ha GH, Nguyen LH, Vu GT, Hoang MT, Le HT, et al. Studies of Novel Coronavirus Disease 19 (COVID-19) pandemic: A Global Analysis of Literature. IJERPH. 2020;17:4095.
26. Aruna S, Rajagopalan S. A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. Int J Comput Appl. 2011;31:14–20.
27. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
29. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017.
30. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–73.
31. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to assess the risk of Bias and Applicability of Prediction Model studies. Ann Intern Med. 2019;170:51.
32. Miceli G, Basso MG, Rizzo G, Pintus C, Cocciola E, Pennacchio AR, et al. Artificial Intelligence in Acute ischemic stroke subtypes according to Toast classification: a Comprehensive Narrative Review. Biomedicines. 2023;11:1138.
33. Wang J, Gong X, Chen H, Zhong W, Chen Y, Zhou Y, et al. Causative classification of ischemic stroke by the machine learning Algorithm Random forests. Front Aging Neurosci. 2022;14:788637.
34. GBD 2016 Neurology Collaborators. 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.
35. Ma Q, Li R, Wang L, Yin P, Wang Y, Yan C, et al. Temporal trend and attributable risk factors of stroke burden in China, 1990–2019: an analysis for the global burden of Disease Study 2019. Lancet Public Health. 2021;6:e897–906.
36. Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. 2010;9:689–701.
37. Xu S, Ilyas I, Little PJ, Li H, Kamato D, Zheng X, et al. Endothelial dysfunction in atherosclerotic Cardiovascular diseases and Beyond: from mechanism to Pharmacotherapies. Pharmacol Rev. 2021;73:924–67.
38. Cai H, Harrison DG. Endothelial dysfunction in cardiovascular diseases: the role of oxidant stress. Circ Res. 2000;87:840–4.
39. Papaharalambus CA, Griendling KK. Basic mechanisms of oxidative stress and reactive oxygen species in cardiovascular injury. Trends Cardiovasc Med. 2007;17:48–54.
40. Wu Z, Yao C, Zhao D, Wu G, Wang W, Liu J, et al. Sino-MONICA project: a collaborative study on trends and determinants in cardiovascular diseases in China, Part I: morbidity and mortality monitoring. Circulation. 2001;103:462–8.
41. Xu G, Ma M, Liu X, Hankey GJ. Is there a stroke belt in China and why? Stroke. 2013;44:1775–83.
42. Li Y, He Y, Lai J, Wang D, Zhang J, Fu P, et al. Dietary patterns are associated with stroke in Chinese adults. J Nutr. 2011;141:1834–9.
43. Liu LS, Tao SC, Lai SH. Relationship between salt excretion and blood pressure in various regions of China. Bull World Health Organ. 1984;62:255–60.
44. Kanhai DA, Kranendonk ME, Uiterwaal CSPM, Van Der Graaf Y, Kappelle LJ, Visseren FLJ. Adiponectin and incident coronary heart disease and stroke. A systematic review and meta-analysis of prospective studies: Adiponectin and risk for future CHD/stroke. Obes Rev. 2013;14:555–67.
45. Straub LG, Scherer PE. Metabolic messengers: Adiponectin. Nat Metab. 2019;1:334–9.
46. Jang AY, Scherer PE, Kim JY, Lim S, Koh KK. Adiponectin and cardiometabolic trait and mortality: where do we go? Cardiovascular Res. 2022;118:2074–84.
47. Chandran M, Phillips SA, Ciaraldi T, Henry RR. Adiponectin: more than just another Fat cell hormone? Diabetes Care. 2003;26:2442–50.
48. Becic T, Studenik C, Hoffmann G. Exercise increases Adiponectin and reduces leptin levels in Prediabetic and Diabetic individuals: systematic review and Meta-analysis of Randomized controlled trials. Med Sci. 2018;6:97.
49. Arsalan null, Ismail M, Khattak MB, Khan F, Anwar MJ, Murtaza Z, et al. Prognostic significance of serum bilirubin in stroke. J Ayub Med Coll Abbottabad. 2011;23:104–7.
50. Lee SJ, Jee YH, Jung KJ, Hong S, Shin ES, Jee SH. Bilirubin and Stroke Risk using a mendelian randomization design. Stroke. 2017;48:1154–60.
51. Ibrahim F, Murr N. Embolic Stroke. StatPearls. Treasure Island (FL). StatPearls Publishing; 2023.
52. Hart RG. Cardiogenic stroke. Am Fam Physician. 1989;40(5 Suppl):S35–8.
53. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al. The causative classification of stroke system: an international reliability and optimization study. Neurology. 2010;75:1277–84.
54. Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. Can one assess whether missing data are missing at random in medical studies? Stat Methods Med Res. 2006;15:213–34.

## Publisher's note

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