Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting 1-Year Recurrence Risk in First-Ever Ischemic Stroke: A Machine Learning Approach Using Routine Clinical Data

Zong Li
Nanchang University
Xia Yang
Nanchang University
Tian Luo
Nanchang University
Jingyi Xu
Nanchang University
Ting Fang
Nanchang University
Zhenglong Huang
Nanchang University
Juan Wan
Nanchang University
Jie Kuang
kuangjie@ncu.edu.cn

Nanchang University

Article

Keywords: First-ever ischemic stroke, Recurrence, Prediction model, Machine learning, SHAP analysis, Routine clinical data

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11504332/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting 1-Year Recurrence Risk in First-Ever Ischemic Stroke: A Machine Learning Approach Using Routine Clinical Data

Zong Li[†], Xia Yang [†], Tian Luo, Jingyi Xu, Ting Fang, Zhenglong Huang, Juan Wan, Jie Kuang[*]

Department of Epidemiology, School of Public Health, Nanchang University, Jiangxi Key Laboratory of Disease Prevention and Public Health, Nanchang 330019, Jiangxi Province, China

[†]Co-first authors

[*]Corresponding author: Kuang Jie, Associate Professor; E-mail: kuangjie@ncu.edu.cn

Abstract

We developed and internally validated machine-learning models to predict 1-year recurrence after first-ever ischemic stroke using routinely available clinical data. This retrospective single-center cohort included patients treated between 2022 and 2024. Predictors were selected using univariate analysis, LASSO regression, and Boruta, and seven machine-learning models were compared with the ESRS and SPI-II scores. Among 1,868 patients, 136 (7.28%) experienced recurrence within 1 year. Nine predictors were retained: age, hyperlipidemia, admission NIHSS score, discharge mRS score, hypertension, diabetes, smoking, LDL-C, and homocysteine. XGBoost showed strong discrimination in the test set (AUC 0.860, 95% CI 0.786–0.933) and significantly outperformed most models, particularly ESRS and SPI-II (both $P<0.001$), whereas differences from LightGBM and random forest were not significant. XGBoost also showed favorable calibration and clinical net benefit. SHAP analysis identified admission NIHSS as the most influential predictor. An R Shiny application was developed for individualized risk estimation. These findings suggest that routinely available clinical variables combined with machine learning may support recurrence risk stratification, although external validation is required before clinical implementation.

Keywords First-ever ischemic stroke, Recurrence, Prediction model, Machine learning, SHAP analysis, Routine clinical data

Introduction

Globally, stroke persists as one of the foremost causes of death and functional disability [1]. Ischemic stroke (IS) accounts for approximately 87% of all strokes [2]. After a first IS, the risk of recurrence is substantial: among patients with transient ischemic attacks or minor strokes, the 1-year recurrence rate is 5.9% and the 10-year rate reaches 19.8% [3]; in young adults, the 1-year recurrence rate ranges from 4.5% to 7.3% [4]. In China, the age-standardized incidence of first-ever stroke is 229.5 per 100,000 [5], and overall stroke incidence increased by 100.6% from 1990 to 2021, with IS showing the largest rise (201.1%) [6]. Recurrent strokes lead to significantly worse outcomes, including higher disability and mortality [7]. Therefore, identifying risk factors for recurrence and developing effective predictive tools are critical for improving patient outcomes.

The Essen Stroke Risk Score (ESRS) and Stroke Prognostic Instrument II (SPI-II) are widely used to stratify the 5-year risk of recurrent stroke and composite vascular events after acute ischemic stroke (AIS). However, their predictive performance is limited in Chinese populations [8, 9], and their longer prediction window (5 years) may not optimally capture the early recurrence risk that is most relevant for prompt secondary prevention. This highlights the need for a tailored prediction model specifically developed for Chinese patients with IS, with a focus on short-term (1-year) recurrence.

Machine learning (ML) has emerged as a powerful tool in medicine, offering new opportunities to advance stroke prevention and risk prediction [10]. ML algorithms can efficiently identify key features associated with clinical outcomes from complex variables, and their predictive performance often surpasses that of conventional statistical methods [11, 12], making them well-suited for improving prognostic accuracy in stroke. Indeed, an increasing number of ML models have been developed for stroke recurrence prediction. A systematic review and meta-analysis of 24,350 patients reported a pooled area under the curve (AUC) of 0.82 for ML models in predicting stroke recurrence [13]. Another large-scale, multicenter study demonstrated that ML models such as XGBoost and LightGBM could effectively predict in-hospital recurrence risk in patients with acute minor ischemic stroke, outperforming traditional clinical scores [14]. Despite these advances, many existing models rely on

specialized or non-routine data, including radiomic features from high-resolution MRI [15], genomics, and gut microbiota [16, 17]. While such models may achieve high accuracy, their dependence on specialized equipment and biomarkers limits their generalizability, particularly in primary and secondary hospitals where only routine clinical data are available. Conversely, models built solely on routine data may sacrifice some predictive accuracy for broader accessibility—a trade-off that is particularly relevant for resource-limited settings.

To address these gaps, this study focused on patients with first-ever IS. Using a three-step feature selection strategy (univariate analysis, LASSO logistic regression, and Boruta algorithm), we identified core risk factors for 1-year recurrence. We then developed and compared seven ML models based solely on routine clinical data, benchmarked the optimal model against the ESRS and SPI-II scores, and built an R Shiny-based real-time prediction platform to provide a proof-of-concept tool for individualized risk assessment that can be further evaluated in prospective clinical settings.

Methods

Study population

This retrospective cohort study consecutively enrolled patients diagnosed with first-ever ischemic stroke (IS) who were hospitalized at the Second Affiliated Hospital of Nanchang University between January 2022 and December 2024. The study was approved by the Biomedical Research Ethics Committee of the Second Affiliated Hospital of Nanchang University (No. 20240522). Inclusion criteria were: (1) age ≥18 years; (2) a first acute IS confirmed by cranial CT/MRI imaging and medical history. Exclusion criteria were: (1) hemorrhagic stroke, mixed stroke, or tumor-related stroke; (2) severe organic dysfunction of the heart, liver, lungs, kidneys, or other major organs; (3) active malignant tumor; (4) in-hospital death.

Data Collection

All baseline data were extracted from the hospital's electronic medical record system, covering the following dimensions: (1) Sociodemographic characteristics: gender, age, etc.; (2) Living habits: smoking, drinking, etc.; (3) Past medical history: hypertension, diabetes, hyperlipidemia, chronic obstructive pulmonary disease, dementia, wake-up stroke, carotid artery stenosis, etc.; (4) Admission clinical signs: systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, respiration, height, weight, body mass index (BMI), National Institutes of Health Stroke Scale (NIHSS) score, etc.; (5) Laboratory test indicators within 24 h after admission: (1) Routine blood test: white blood cell count, red blood cell count, platelet count, lymphocyte count, neutrophil count, etc.; (2) Biochemical indicators: urea, creatinine, α-hydroxybutyric dehydrogenase, total protein, triglyceride, apolipoprotein A, high-density lipoprotein cholesterol, blood glucose, homocysteine, etc.; (3) Prothrombin time, prothrombin activity.

Definitions of relevant variables: (1) Smoking/drinking history: refers to the relevant behavior that persisted or accumulated for ≥6 months in the past; (2) Hypertension: previously diagnosed, taking antihypertensive drugs, or admission SBP ≥140 mmHg/DBP ≥90 mmHg; (3) Diabetes: previously diagnosed, using hypoglycemic drugs, or fasting blood glucose ≥7.0 mmol/L or random blood glucose ≥11.1 mmol/L; (4) Hyperlipidemia: previously diagnosed, or meeting any of the following criteria: total cholesterol ≥6.2 mmol/L, triglyceride ≥2.3 mmol/L,

low-density lipoprotein cholesterol (LDL-C) ≥4.1 mmol/L, high-density lipoprotein cholesterol <1.0 mmol/L; (5) Wake-up stroke: an ischemic stroke presenting with neurological deficits upon awakening, where the exact time of symptom onset cannot be determined.

Outcome Definition and Follow-up

The primary outcome was ischemic stroke recurrence within one year after discharge, defined as readmission due to new neurological deficits with confirmation of a new cerebral infarct on cranial CT/MRI. Follow-up was conducted by two trained neurologists via telephone or outpatient visits. The follow-up period began on the discharge date and ended at 1 year post-discharge or on the date of first confirmed recurrence, whichever occurred first.

Data Preprocessing and Variable Selection

Before model construction, rigorous data quality control was implemented. First, we analyzed the pattern of missing data. To minimize noise, variables with a missing rate exceeding 20% were excluded from the analysis. For the remaining variables, missing values were multiply imputed using the Multivariate Imputation by Chained Equations (MICE) algorithm All continuous features were standardized using the Z-score method, and categorical variables were one-hot encoded. A three-step approach was used to select core predictive variables, with statistical significance set at $P < 0.05$ for each step: (1) Univariate analysis: Baseline characteristics were compared between the recurrence and non-recurrence groups using the *t*-test, Mann-Whitney *U* test, chi-square test, or Fisher's exact test, as appropriate. Variables showing significant differences were retained. (2) LASSO logistic regression: Variables significant in univariate analysis were further analyzed using LASSO logistic regression. The optimal regularization parameter λ was selected via five-fold cross-validation combined with grid search ($\lambda \in [0.001, 0.01, 0.1, 1, 10]$), and variables with non-zero coefficients were retained. (3) Boruta algorithm: The Boruta algorithm was applied to the variables significant in univariate analysis, and important predictors were identified by comparing feature Z-scores with those of shadow features. Variables identified by all three methods were defined as the core feature set for subsequent model development.

Machine Learning Model Development

The overall process of data preprocessing and model development in this study is illustrated in

Figure 1. The original dataset contained 10 variables, including the outcome variable (Relapse) and 9 potential predictors. First, variable renaming and type conversion were performed: ordinal variables were converted to numeric, the outcome variable was converted to a binary factor, and categorical variables were converted to factor type. Subsequently, the data were divided into a training set (70%) and a test set (30%) using stratified random sampling, ensuring that the distribution of the outcome variable was consistent between the two sets.

To address class imbalance in the training set, class weights were calculated (with the minority class receiving higher weight) and applied in all subsequent model training to improve the model's ability to identify the minority class.

A total of seven machine learning models were constructed: logistic regression, XGBoost, random forest, neural network, support vector machine, K-nearest neighbors, and LightGBM. The hyperparameters of each model were optimized based on the training set using cross-validation or early stopping strategies: logistic regression employed 10-fold cross-validation to select the regularization strength of the elastic net; XGBoost and LightGBM used validation-set early stopping to determine the optimal number of boosting rounds, supplemented by shallow tree structures, low learning rates, and regularization parameters to prevent overfitting; random forest reduced variance through balanced sampling and by limiting tree depth and node size; the neural network adopted a very small hidden layer structure with weight decay; the support vector machine used a linear kernel with adjusted class weights; K-nearest neighbors applied feature scaling and selected a relatively large K value to improve generalization.

After fitting all models on the training set, the optimal classification threshold for each model was determined by maximizing the Youden index based on the predicted probabilities from the training set. Model performance was comprehensively evaluated using the following metrics: area under the receiver operating characteristic curve (AUC) with its 95% confidence interval, accuracy, sensitivity, specificity, precision, F1-score, Youden index, and balanced accuracy. Finally, the model that achieved the best performance on the majority of metrics on the test set was selected as the final prediction model.

Model Interpretability

SHapley Additive exPlanations (SHAP) analysis was performed to interpret the predictive mechanisms of the optimal XGBoost model, quantifying the contribution of each feature to individual predictions [18, 19]. Finally, a real-time interactive prediction platform was developed using the R Shiny framework, enabling instantaneous recurrence risk estimation to support clinical decision-making.

Comparison with Conventional Risk Stratification Tools

To enable comparison with established clinical risk scores, a modified clinical risk score was constructed based on the core design principles of the Essen Stroke Risk Score (ESRS) and Stroke Prognostic Instrument II (SPI-II). For ESRS, a complete subset of variables available in this dataset was used, including age, history of hypertension, history of diabetes, smoking history, and peripheral arterial disease. For SPI-II, the extracted variables included age, Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification, SBP, and history of diabetes.

ESRS and SPI-II scores were calculated by two trained neurologists within 7 days of admission. Inter-rater reliability was assessed using the Kappa statistic (Kappa = 0.89, indicating good agreement). Patients were stratified according to ESRS scores into low-risk (0–2 points) and high-risk (3–6 points) groups, and according to SPI-II scores into low-risk (0–3 points), moderate-risk (4–7 points), and high-risk (8–12 points) groups. The scoring methods are detailed in Table S1.

Statistical Analysis

Normally distributed continuous data were expressed as mean (SD) and compared by t-test; non-normal continuous data were presented as median (IQR) and analyzed using the Mann-Whitney U test. Categorical data were shown as frequencies with percentages and tested using either the chi-square test or Fisher's exact test as appropriate. A two-sided $P < 0.05$ denoted statistical significance, and 95% confidence intervals (CI) were given where applicable. All statistical analyses and visualizations were conducted using R version 4.5.2 (R Foundation for Statistical Computing), with the pROC, ggplot2, shapviz, shiny, mice, and caret packages.

Results

Baseline Characteristics of Patients

A total of 1868 patients with first-ever ischemic stroke (IS) were enrolled, of whom 136 (7.28%) experienced recurrence within one year. Significant differences in baseline characteristics were observed between the recurrence and non-recurrence groups (Table 1). Patients in the recurrence group were older and had significantly higher proportions of hyperlipidemia, hypertension, diabetes, admission National Institutes of Health Stroke Scale (NIHSS) score ≥16, and elevated low-density lipoprotein cholesterol (LDL-C) levels compared with the non-recurrence group (all $P < 0.05$).

Screening of Predictive Variables

Fifteen variables with statistical significance in univariate analysis were included in the subsequent screening process. LASSO regression analysis identified the optimal regularization parameter via five-fold cross-validation using the one-standard-error rule (λ.1se = 0.0095; Figures 2A and 2B), yielding nine features with non-zero coefficients: age, history of hyperlipidemia, history of diabetes, history of hypertension, admission NIHSS score, discharge modified Rankin Scale (mRS) score, smoking history, homocysteine (HCY) concentration, and LDL-C level.

The Boruta algorithm identified 11 important predictors: age, history of hyperlipidemia, albumin, albumin/globulin ratio, history of diabetes, history of hypertension, admission NIHSS score, discharge mRS score, smoking history, HCY concentration, and LDL-C level (Figure 2C). The intersection of the two screening strategies yielded nine consistent predictive factors (Figure 2D), which were defined as the core feature set for the final prediction model.

Comparison of Prediction Model Performance

ROC curves for the seven ML models and two conventional stroke risk scores are shown in Figure 3. In the training set, XGBoost achieved the highest area under the ROC curve (AUC) of 0.920 (95% CI: 0.892–0.949), followed by LightGBM (AUC = 0.920, 95% CI: 0.892–0.949) and random forest (RF) (AUC = 0.882, 95% CI: 0.843–0.920). Among the traditional models, the AUC of SPI-II was 0.627 (95% CI: 0.572–0.682), and that of ESRS was only 0.558 (95% CI: 0.497–0.612) (Figures 3A and 3C; Table 2). In the test set, XGBoost maintained optimal

discriminative performance (AUC = 0.860, 95% CI: 0.786–0.933), followed by LightGBM (AUC = 0.860, 95% CI: 0.786–0.933) and RF (AUC = 0.835, 95% CI: 0.763–0.907); ESRS (0.523, 95% CI: 0.424–0.625) and SPI-II (0.525, 95% CI: 0.427–0.622) had similar and significantly lower performance (Figures 3B and 3D).

The DeLong test indicated that the performance of XGBoost was significantly better than that of most other models, particularly the SPI-II and ESRS models (p=0.000), whereas no statistically significant differences were observed between XGBoost and the suboptimal models, namely LightGBM and random forest (RF). This finding indicates that several high-performance machine learning models possess comparable discriminative ability. Consequently, XGBoost was ultimately selected as the optimal model, achieving the highest accuracy (0.879), lowest log loss (0.152), and highest specificity (0.888), precision (0.352), and F1-score (0.483) in the training set. Calibration analysis demonstrated good agreement between predicted and observed recurrence probabilities for XGBoost (Figures 4A and 4B). Decision curve analysis further confirmed the clinical utility of XGBoost, which provided the greatest net benefit across all threshold probability ranges (Figures 4C and 4D).

Feature Importance Analysis

The mean absolute SHAP values were calculated to assess the relative importance of core features in the XGBoost model (Figure 5A). Admission NIHSS score was the most influential predictor, followed by history of hypertension, age, and HCY concentration. The contribution of admission NIHSS score substantially exceeded that of other features, underscoring its central role in recurrence prediction.

SHAP waterfall plots were generated for true positive (recurrent) and true negative (non-recurrent) cases (Figures 5C and 5D). In true positive cases, admission NIHSS score contributed the largest positive predictive value (+0.341), whereas discharge mRS score contributed a negative value. In contrast, admission NIHSS score showed a markedly lower or negative contribution in true negative cases, and the cumulative contribution of key risk features did not reach the classification threshold. These findings indicate that the model effectively captured the distinct feature profiles differentiating recurrent from non-recurrent patients.

Clinically Implementable Application

To translate the XGBoost prediction model into clinical practice, an integrated real-time prediction platform was developed using the R Shiny package (Figure 6). The platform comprises three dynamically linked core modules: (1) Model overview panel: Displays the hyperparameters, performance metrics, and core feature set of the optimal XGBoost model. (2) SHAP interpretation module: Provides interactive feature importance bar charts and dependence swarm plots to visualize global model decision patterns and individual feature contributions. (3) Real-time prediction interface: Allows clinicians to input patient clinical features and obtain immediate recurrence risk estimates, including recurrence probability and risk level, accompanied by intuitive risk interpretation.

This platform integrates “precise prediction” with “individualized interpretation,” offering a practical tool to enhance clinical decision-making for stroke recurrence prevention. The final model is presented as an interactive web application built using the R package Shiny , which is accessible at the following URL: https://cardiowise.shinyapps.io/stroke-predictor/. A detailed illustrative case example is provided in Figure 6. Briefly, for a 65-year-old patient with a history of hypertension and diabetes, admission NIHSS score of 13 (moderate stroke), discharge mRS score of 3, LDL-C of 3.5 mmol/L, HCY of 15.0 μmol/L, and a smoking history, the model predicts a recurrence probability of 0.395, which exceeds the threshold of 0.082 and classifies the patient as high-risk. In contrast, for a 55-year-old patient without any of these risk factors and an admission NIHSS score of 2, the predicted probability is 0.019, classifying the patient as low-risk. These examples are for illustrative purposes only and should not replace clinical judgment.

Discussion

In this study, we followed 1868 patients with first-ever IS for one year, identified nine core recurrence risk factors using a three-step screening strategy combining univariate analysis, and constructed seven ML models. The XGBoost model demonstrated the best predictive performance, and we further developed an R Shiny-based real-time prediction platform to facilitate clinical translation. Our findings provide a locally adapted prediction tool and clear modifiable intervention targets for secondary prevention of first-ever IS recurrence in Chinese patients.

The nine core predictive variables identified in this study span sociodemographic characteristics, comorbidities, clinical functional scores, and laboratory biomarkers, collectively capturing the multifaceted clinical profile of Chinese patients with first-ever IS. SHAP analysis revealed that the admission National Institutes of Health Stroke Scale (NIHSS) score was the most important predictor of recurrence [18, 19], a finding consistent with international stroke prognosis studies [20, 21]. The NIHSS score directly quantifies the severity of initial neurological deficits. In our cohort, the proportion of patients with severe stroke (NIHSS ≥16) was 65.4% in the recurrence group, markedly higher than the 20.9% in the non-recurrence group. This suggests that large infarct size, involvement of critical cerebrovascular territories, and insufficient collateral circulation constitute the core pathological substrates for IS recurrence [20]. The prognostic value of NIHSS has also been validated in studies on basilar artery occlusion [21], further supporting its role as a key predictive indicator. Moreover, a recent nationwide propensity-matched study of 39,947 first-ever IS patients reported that discharge NIHSS score significantly predicts early recurrence, with patients having NIHSS ≥5 experiencing 20%–30% higher recurrence rates, further underscoring the prognostic importance of neurological deficit severity throughout the acute phase and early recovery period [22].

Homocysteine (HCY) increases recurrence risk by inducing oxidative damage to the vascular endothelium and enhancing thrombotic tendency [23], while low-density lipoprotein cholesterol (LDL-C)—a central pathogenic factor in atherosclerosis—directly promotes the formation and rupture of atherosclerotic plaques [24]. A 2025 review further elucidated that

HCY metabolites promote endothelial dysfunction through epigenetic mechanisms involving specific microRNAs and histone modifications, thereby accelerating atherosclerosis and increasing stroke risk [25]. In addition, diabetes and smoking further elevate recurrence risk by inducing a prothrombotic state and persistent endothelial injury [26, 27]. In our study, the prevalence of diabetes in the recurrence group was 39.7%, significantly higher than in the non-recurrence group, aligning with findings from a meta-analysis on the impact of diabetes on stroke outcomes [26]. The identification of these modifiable risk factors provides clear targets for individualized secondary prevention strategies in clinical practice.

All ML models developed in this study significantly outperformed the traditional ESRS and SPI-II scales (test set AUC = 0.523 and 0.525, respectively), with XGBoost achieving the best performance (test set AUC = 0.860). This finding underscores the unique advantages of ML in addressing complex multifactorial prediction problems such as IS recurrence [11, 12]. This superiority is further supported by a recent large-scale, multicenter Chinese study of 322,135 patients with acute minor ischemic stroke, which reported that ML models (XGBoost and LightGBM) achieved AUCs of 0.788–0.803, significantly outperforming traditional generalized linear models [14]. Moreover, the DISCO modelling study based on the China National Stroke Registry (CNSR-III) demonstrated that Δ-NIHSS (admission-to-discharge) was the strongest predictor of stroke recurrence, with an AUC of 0.805 for 3-month recurrence and a 19.27-fold relative risk in the highest-risk 1% of patients [28]. The limited predictive ability of conventional scales can be attributed to three key limitations: first, they were developed in Western populations and do not account for clinical characteristics of Chinese patients, such as the high prevalence of hypertension and hyperhomocysteinemia [8, 9]; second, their linear scoring frameworks cannot capture the complex nonlinear relationships and interactions among variables that drive recurrence [29]; third, they include a restricted set of variables and do not incorporate biomarkers or functional scores with strong predictive value, such as NIHSS, modified Rankin Scale (mRS), LDL-C, and HCY [8, 9]. A 2026 systematic review and meta-analysis of 14 ML studies in AIS confirmed the robust performance of ML models, with a pooled AUC of 0.87, while also noting that substantial heterogeneity across studies underscores the need for standardized development and validation protocols to enhance

model generalizability [30].

As a representative ensemble learning algorithm, XGBoost efficiently captures nonlinear and interaction effects through gradient-boosted decision trees [31, 32]. Its built-in regularization mechanism mitigates overfitting [33], making it well-suited for heterogeneous clinical data in IS. These advantages have also been demonstrated in studies predicting hospital length of stay in stroke patients [33]. Furthermore, we applied SHAP analysis to interpret the XGBoost model, quantifying each factor‘s predictive contribution [18]. This approach addresses the “black box” limitation of traditional ML models [34], enhancing the interpretability and clinical acceptability of predictions—a critical step toward clinical translation.

The core value of this study lies in two key aspects: methodological rigor and clinical implementability. First, we employed a three-step feature selection strategy integrating univariate analysis, LASSO logistic regression, and the Boruta algorithm. This approach mitigates the bias inherent in any single screening method, ensures the reliability and stability of the core variables, and provides a replicable methodological framework for prognostic research in cerebrovascular diseases [35]. Second, we developed a locally adapted prediction model for first-ever IS patients in southern China. The identified core risk factors align with the clinical characteristics of the Chinese population [8, 9], addressing the poor applicability of Western-derived scales and foreign ML models. Third, we constructed an R Shiny platform that integrates model overview, SHAP interpretation, and real-time prediction [36], enabling the integration of “precise prediction” with “individualized interpretation.” The use of R Shiny for clinical risk prediction has been increasingly adopted; for instance, a recent study developed a dynamic nomogram for prehospital delay in first-episode AIS using R Shiny, achieving an AUC of 0.847 in the training set and 0.834 in the validation set, demonstrating the platform’s utility and scalability for real-time clinical decision support [37]. This tool addresses the gap in previous ML studies that focused primarily on model development without providing practical clinical translation solutions [36]. The platform is user-friendly and yields intuitive results, making it suitable for application in both tertiary hospitals and primary care settings.

Several limitations of this study need to be acknowledged. First, the data were derived from a single-center retrospective cohort, with most participants recruited from Nanchang, Jiangxi

Province. This geographic and demographic concentration may restrict the applicability of the developed model to broader Chinese populations or other ethnic groups. Second, our predictors were limited to routine clinical, functional, and biochemical variables. Emerging biomarkers linked to stroke recurrence, including radiomic features of intracranial atherosclerotic plaques, genetic polymorphisms, and gut microbiota composition, were not incorporated. Integrating such multi-source data might further improve model accuracy. Third, only internal validation was performed; no external validation using independent cohorts from other hospitals or regions was carried out. Additionally, our model is static and fails to account for time-varying changes in key risk factors (e.g., blood pressure and glucose levels) during follow-up, highlighting the need for a dynamic prediction framework based on longitudinal data. Fourth, the R Shiny platform has not yet undergone large-scale clinical trials or health economic assessments. Its performance in real-world settings, acceptance by clinicians, and cost-effectiveness remain to be investigated.

Conclusion

In summary, this study identified nine core predictors—including age, history of hyperlipidemia, and admission NIHSS score—that substantially enhance the accuracy of predicting 1-year recurrence risk in patients with first-ever ischemic stroke. Among the evaluated methods, the XGBoost algorithm outperformed other machine learning approaches and conventional risk scoring systems in both predictive performance and clinical applicability, with the admission NIHSS score standing out as a particularly influential factor. The interactive prediction tool built on the R Shiny framework offers an intuitive and interpretable interface to facilitate personalized clinical decision-making regarding stroke recurrence.

List of abbreviations

abbreviation	full name
XGBoost	Extreme Gradient Boosting
NN	Neural Network
RF	Random Forest
LR	Logistic Regression
LightGBM	Light Gradient Boosting Machine
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
SPI-II	Stroke Prognostic Instrument II
ESRS	Essen Stroke Risk Score
AUC	area under the curve
DCA	decision curve analysis
SHAP	Shapley Additive Explanations
TIA	transient ischemic attack
NIHSS	National Institutes of Health Stroke Scale
mRS	modified Rankin Scale
COPD	chronic obstructive pulmonary disease
DBP	diastolic blood pressure
BMI	body mass index
WBC	white blood cell count
RBC	red blood cell count
Hb	hemoglobin
PLT	platelet count
Hct	hematocrit
MCV	mean corpuscular volume
MCH	mean corpuscular hemoglobin
MCHC	mean corpuscular hemoglobin concentration
Pct	plateletcrit
MPV	mean platelet volume
PLCR	platelet large cell ratio
LYM	lymphocyte count
MONO	monocyte count
NEUT	neutrophil count
TP	total protein

ALB	albumin
GLB	globulin
A/G ratio	albumin/globulin ratio
TBIL	total bilirubin
DBIL	direct bilirubin
IBIL	indirect bilirubin
AST	aspartate aminotransferase
ALT	alanine aminotransferase
ALP	alkaline phosphatase
Cr	creatinine
UA	uric acid
CK	creatine kinase
CK-MB	creatine kinase-MB isoenzyme
LDH	lactate dehydrogenase
α-HBDH	α-hydroxybutyric dehydrogenase
TC	total cholesterol
TG	triglycerides
HDL-C	high-density lipoprotein cholesterol
LDL-C	low-density lipoprotein cholesterol
Non-HDL-C	non-high-density lipoprotein cholesterol
ApoA	apolipoprotein A
ApoB	apolipoprotein B
Lp(a)	lipoprotein(a)
PT	prothrombin time
APTT	activated partial thromboplastin time
TT	thrombin time
BG	blood glucose
HCY	homocysteine

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

Funding: National Natural Science Foundation of China (82160645, 82360667); Jiangxi Provincial Natural Science Foundation (20212BAB206091); National College Students' Innovation and Entrepreneurship Training Program (202510403077)

Ethical approval and consent to participate

This study was conducted in compliance with the Declaration of Helsinki and approved by the Biomedical Research Ethics Committee of the Second Affiliated Hospital of Nanchang University (No. 20240522). All data was anonymized and the informed consent was waived by our IRB due to the study's retrospective nature.

Acknowledgement

We would like to thank the colleagues for helping data collection.

Author contributions

Zong Li: Conceptualization, data curation, formal analysis, methodology, software, visualization, and writing—original draft. Xia Yang: Conceptualization, data curation, formal analysis, investigation, validation, supervision, and writing—review and editing. Tian Luo: Investigation, data curation, and formal analysis. Jingyi Xu: Investigation and software. Ting Fang: Data curation and visualization. Zhenglong Huang: Methodology and validation. Juan Wan: Resources and supervision. Jie Kuang: Conceptualization, methodology, supervision, validation, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Additional information

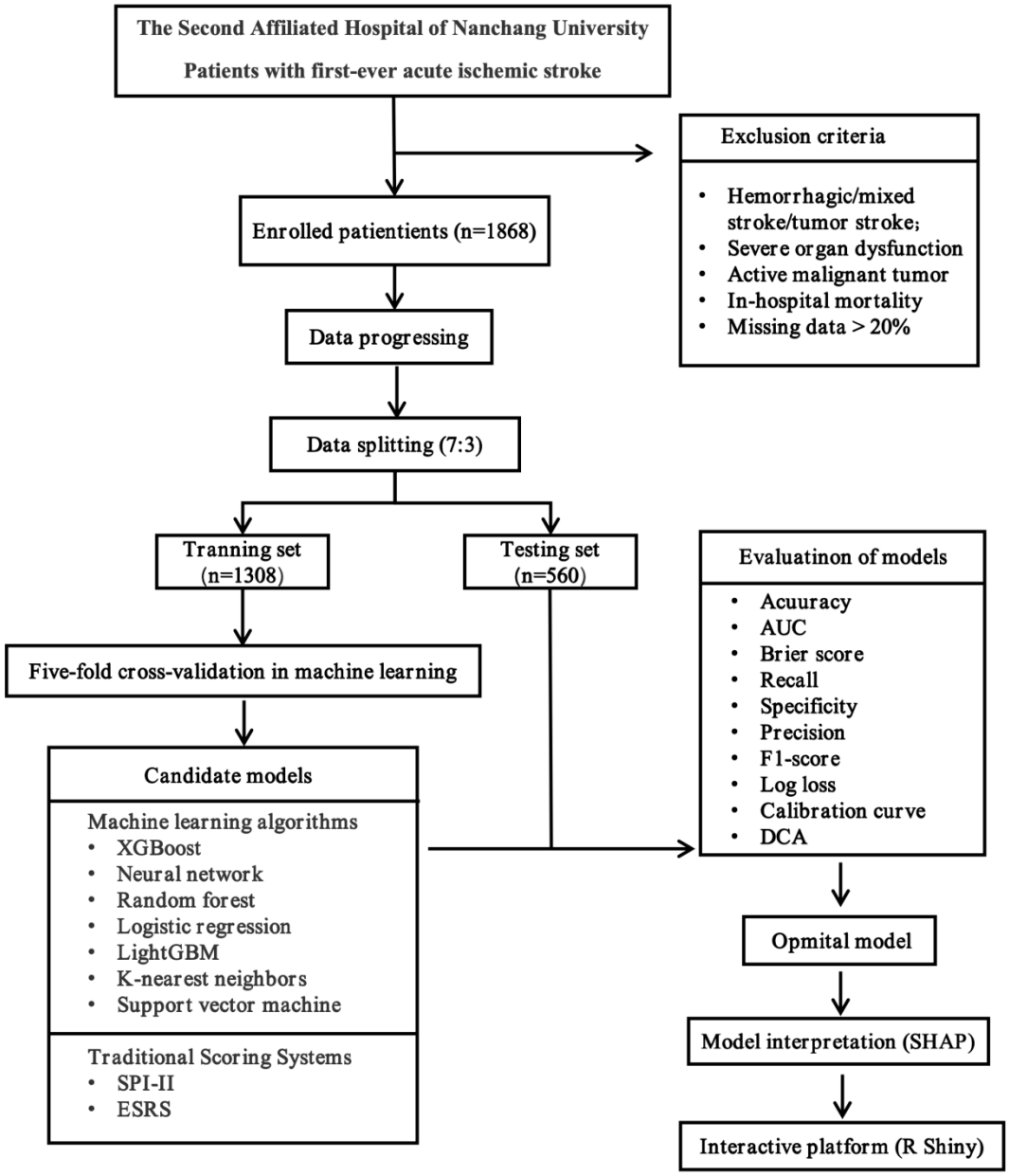
Supplementary material: Supplementary material is provided with this submission (file: Supplementary materials.docx), containing: Supplementary Tables S1–S4 and Supplementary Fig. S1.

References

1. Martin, S. S. et al. 2025 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 151, e41–e660. https://doi.org/10.1161/CIR.0000000000001303 (2025).
2. Zhao, H. et al. Comprehensive analysis of the ischemic stroke burden at global, regional, and national levels (1990–2021): trends, influencing factors, and future projections. Front. Neurol. 16, 1492691. https://doi.org/10.3389/fneur.2025.1492691 (2025).
3. Writing Committee for the PERSIST Collaborators. Long-term risk of stroke after transient ischemic attack or minor stroke: a systematic review and meta-analysis. JAMA 333, 1508–1519. https://doi.org/10.1001/jama.2025.2033 (2025).
4. Kwok, G. Y. R. et al. Recurrent ischemic stroke in young adults: a multicenter cohort study, systematic review, and meta-analysis. Int. J. Stroke 21, 24–35. https://doi.org/10.1177/17474930251340799 (2026).
5. Sun, H. et al. Incidence of first-ever stroke, stroke events, and early case fatality rate in China: results from a national population-based survey. Neuroepidemiology 59, 714–726. https://doi.org/10.1159/000543474 (2025).
6. Wang, J. et al. Analysis and prediction of disease burden of stroke and its subtypes in China from 1990 to 2040. Chin. Med. J. (Engl.) 138, 2452–2463. https://doi.org/10.1097/CM9.0000000000003725 (2025).
7. Jiang, Q. et al. Decade trends in risk factors, treatments, and prognosis of minor ischemic stroke and TIA in China. Int. J. Stroke 20, 1090–1101. https://doi.org/10.1177/17474930251342357 (2025).
8. Zhao, J., Wang, D., Liu, X., Wang, Y. & Zhao, X. The predictive value of Essen and SPI-II on the risk of 5-year recurrence in Chinese patients with acute ischemic stroke. Neuropsychiatr. Dis. Treat. 19, 2251–2260. https://doi.org/10.2147/NDT.S433383 (2023).
9. Kongwatcharapong, J., Sornkhamphan, A., Kaveeta, C. & Nathisuwan, S. Validation and comparison of the stroke prognosis instrument (SPI-II) and the Essen stroke risk score (ESRS) in predicting stroke recurrence in Asian population. BMC Neurol. 23, 287. https://doi.org/10.1186/s12883-023-03329-w (2023).
10. Liu, J., Li, M. & Liu, J. Evaluation of the ESRS and SPI-II scales for short-term prognosis of minor stroke and transient ischemic attack. Neurol. Res. 35, 568–572. https://doi.org/10.1179/1743132813Y.0000000172 (2013).
11. Han, H. & Liu, W. The coming era of artificial intelligence in biological data science. BMC Bioinformatics 20, 712. https://doi.org/10.1186/s12859-019-3225-3 (2019).
12. Heo, J. N. et al. Machine learning-based model for prediction of outcomes in acute stroke. Stroke 50, 1263–1265. https://doi.org/10.1161/STROKEAHA.118.024293 (2019).
13. Habibi, M. A. et al. The performance of machine learning for predicting the recurrent stroke: a systematic review and meta-analysis on 24,350 patients. Acta Neurol. Belg. 125, 609–624. https://doi.org/10.1007/s13760-024-02682-y (2025).
14. Ye, W. et al. A combined-mode machine learning model for predicting stroke recurrence during hospitalization in patients with acute minor ischemic stroke. MedComm 6,

e70234. https://doi.org/10.1002/mco2.70234 (2025).
15. Li, Z. et al. Predicting the prognosis of symptomatic intracranial atherosclerotic stenosis (sICAS) patients using deep learning models: a multicenter study based on high-resolution magnetic resonance vessel wall imaging. Clin. Radiol. 91, 107092. https://doi.org/10.1016/j.crad.2025.107092 (2025).
16. Chauhan, G. & Debette, S. Genetic risk factors for ischemic and hemorrhagic stroke. Curr. Cardiol. Rep. 18, 124. https://doi.org/10.1007/s11886-016-0804-z (2016).
17. Peh, A., O'Donnell, J. A., Broughton, B. R. S. & Marques, F. Z. Gut microbiota and their metabolites in stroke: a double-edged sword. Stroke 53, 1788–1801. https://doi.org/10.1161/STROKEAHA.121.036800 (2022).
18. Nohara, Y., Matsumoto, K., Soejima, H. & Nakashima, N. Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. Comput. Methods Programs Biomed. 214, 106584. https://doi.org/10.1016/j.cmpb.2021.106584 (2022).
19. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9 (2020).
20. Alemseged, F. et al. Posterior National Institutes of Health Stroke Scale improves prognostic accuracy in posterior circulation stroke. Stroke 53, 1247–1255. https://doi.org/10.1161/STROKEAHA.120.034019 (2022).
21. Lu, V. M. et al. Presenting NIHSS predicts 90-day functional outcome after mechanical thrombectomy for basilar artery occlusion: a systematic review and meta-analysis. Clin. Neurol. Neurosurg. 197, 106199. https://doi.org/10.1016/j.clineuro.2020.106199 (2020).
22. Kim, Y., Jeong, H. S., Shim, S. R. & Hyun, J. K. Discharge neurological deficit as a predictor of early stroke recurrence: a nationwide registry-based propensity-matched study. J. Am. Heart Assoc. 14, e043529. https://doi.org/10.1161/JAHA.125.043529 (2025).
23. Chrysant, S. G. & Chrysant, G. S. The current status of homocysteine as a risk factor for cardiovascular disease: a mini review. Expert Rev. Cardiovasc. Ther. 16, 559–565. https://doi.org/10.1080/14779072.2018.1497974 (2018).
24. Wilson, P. W. F. et al. Systematic review for the 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol. J. Am. Coll. Cardiol. 73, 3210–3227. https://doi.org/10.1016/j.jacc.2018.11.004 (2019).
25. Jakubowski, H. & Witucki, Ł. Homocysteine metabolites, endothelial dysfunction, and cardiovascular disease. Int. J. Mol. Sci. 26, 746. https://doi.org/10.3390/ijms26020746 (2025).
26. Lau, L.-H., Lew, J., Borschmann, K., Thijs, V. & Ekinci, E. I. Prevalence of diabetes and its effects on stroke outcomes: a meta-analysis and literature review. J. Diabetes Investig. 10, 780–792. https://doi.org/10.1111/jdi.12932 (2019).
27. Chen, J. et al. Impact of smoking status on stroke recurrence. J. Am. Heart Assoc. 8, e011696. https://doi.org/10.1161/JAHA.118.011696 (2019).
28. Cheng, S. et al. Artificial intelligence-powered delta-NIHSS-based model for predicting recurrence, disability and mortality after acute ischaemic strokes (DISCO): a modelling

study. eClinicalMedicine 90, 103684. https://doi.org/10.1016/j.eclinm.2025.103684 (2025).
29. Chaudhary, D. et al. Clinical risk score for predicting recurrence following a cerebral ischemic event. Front. Neurol. 10, 1106. https://doi.org/10.3389/fneur.2019.01106 (2019).
30. Khanum, U., Guddattu, V., Paneyala, S., Srinivasan, A. & Nagaraju, C. Predictive performance of machine learning models in acute ischemic stroke: a systematic review and meta-analysis. Front. Neurol. 17, 1771341. https://doi.org/10.3389/fneur.2026.1771341 (2026).
31. Ke, L. et al. Risk factors and prediction models for recurrent acute ischemic stroke: a retrospective analysis. PeerJ 12, e18605. https://doi.org/10.7717/peerj.18605 (2024).
32. Li, J. et al. Tree-based risk factor identification and stroke level prediction in stroke cohort study. Biomed Res. Int. 2023, 7352191. https://doi.org/10.1155/2023/7352191 (2023).
33. Chen, R. et al. A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm. BMC Med. Inform. Decis. Mak. 23, 49. https://doi.org/10.1186/s12911-023-02140-4 (2023).
34. Lv, J. et al. An interpretable machine learning approach for predicting 30-day readmission after stroke. Int. J. Med. Inform. 174, 105050. https://doi.org/10.1016/j.ijmedinf.2023.105050 (2023).
35. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. Front. Bioinform. 2, 927312. https://doi.org/10.3389/fbinf.2022.927312 (2022).
36. Feng, G. et al. Artificial intelligence in chronic disease management for aging populations: a systematic review of machine learning and NLP applications. Int. J. Gen. Med. 18, 3105–3115. https://doi.org/10.2147/IJGM.S516247 (2025).
37. Li, F., Zhang, L., Zhang, R., Liu, Y. & Zhang, T. Development and validation of a dynamic nomogram of prehospital delay in patients with first-episode acute ischemic stroke. Geriatr. Nurs. 66, 103632. https://doi.org/10.1016/j.gerinurse.2025.103632 (2025).



519

520 **Figure 1.** Overview of the study design and machine learning workflow. Abbreviations:

521 XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SPI-II,

522 Stroke Prognostic Instrument II; ESRS, Essen Stroke Risk Score; AUC, area under the curve;

523 DCA, decision curve analysis; SHAP, Shapley Additive Explanations.

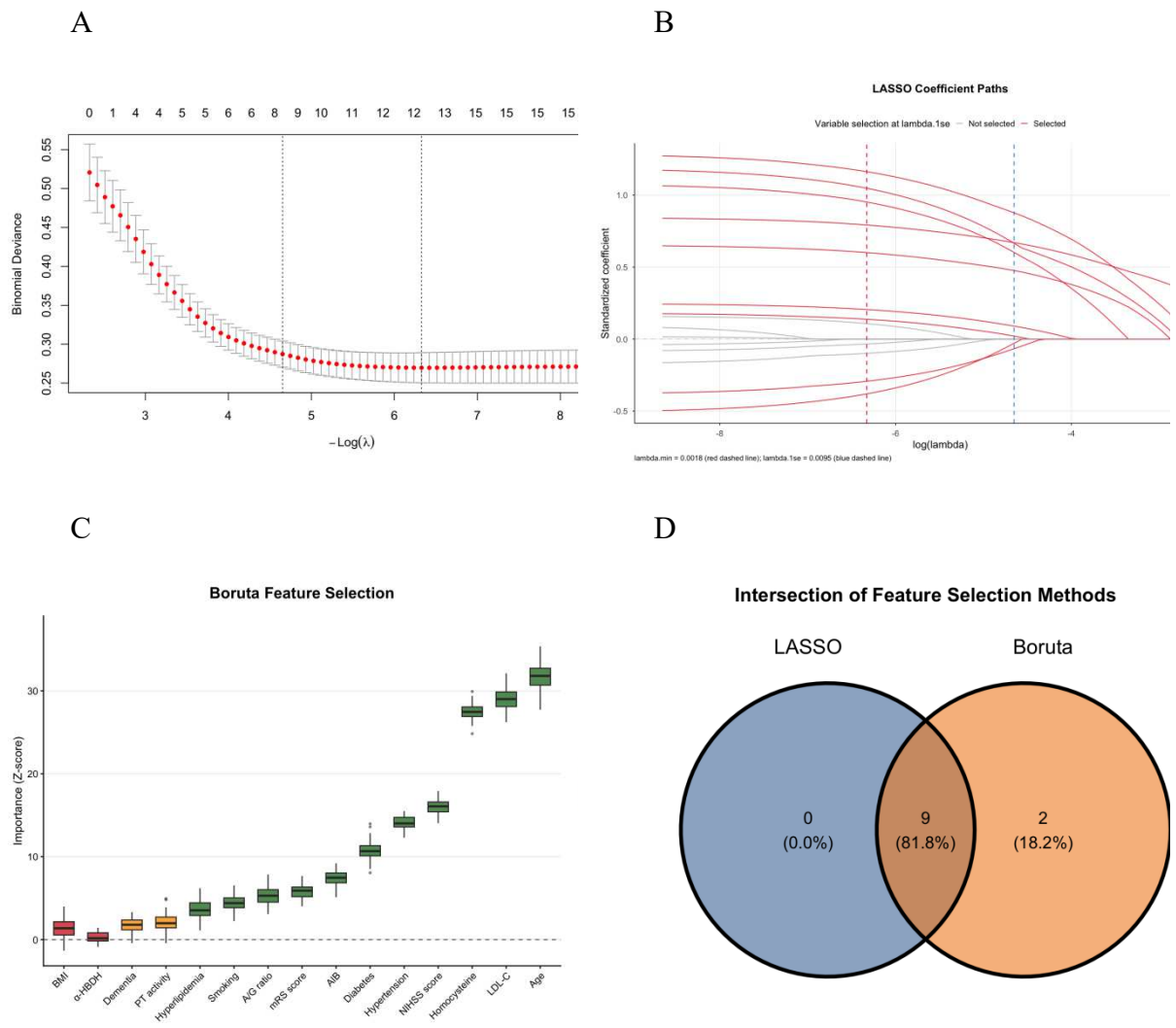


Figure 2. Feature variable screening using LASSO regression and the Boruta algorithm. (A) Coefficient curves for clinical features in LASSO regression; (B) Screening of the optimal regularization parameter by 5-fold cross-validation in LASSO regression; (C) Boruta feature importance plot; (D) Venn diagram of feature selection by LASSO regression and the Boruta algorithm.

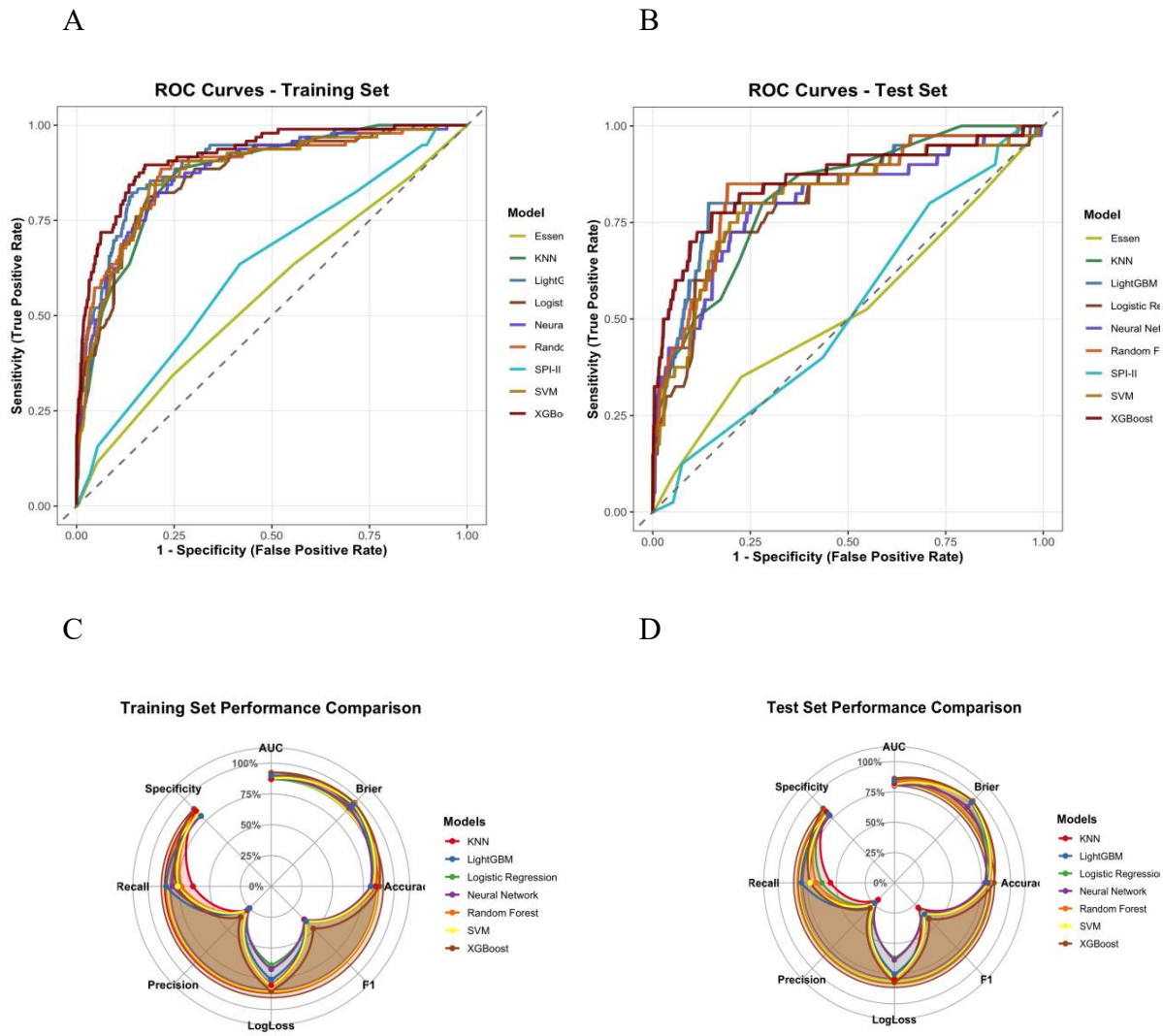
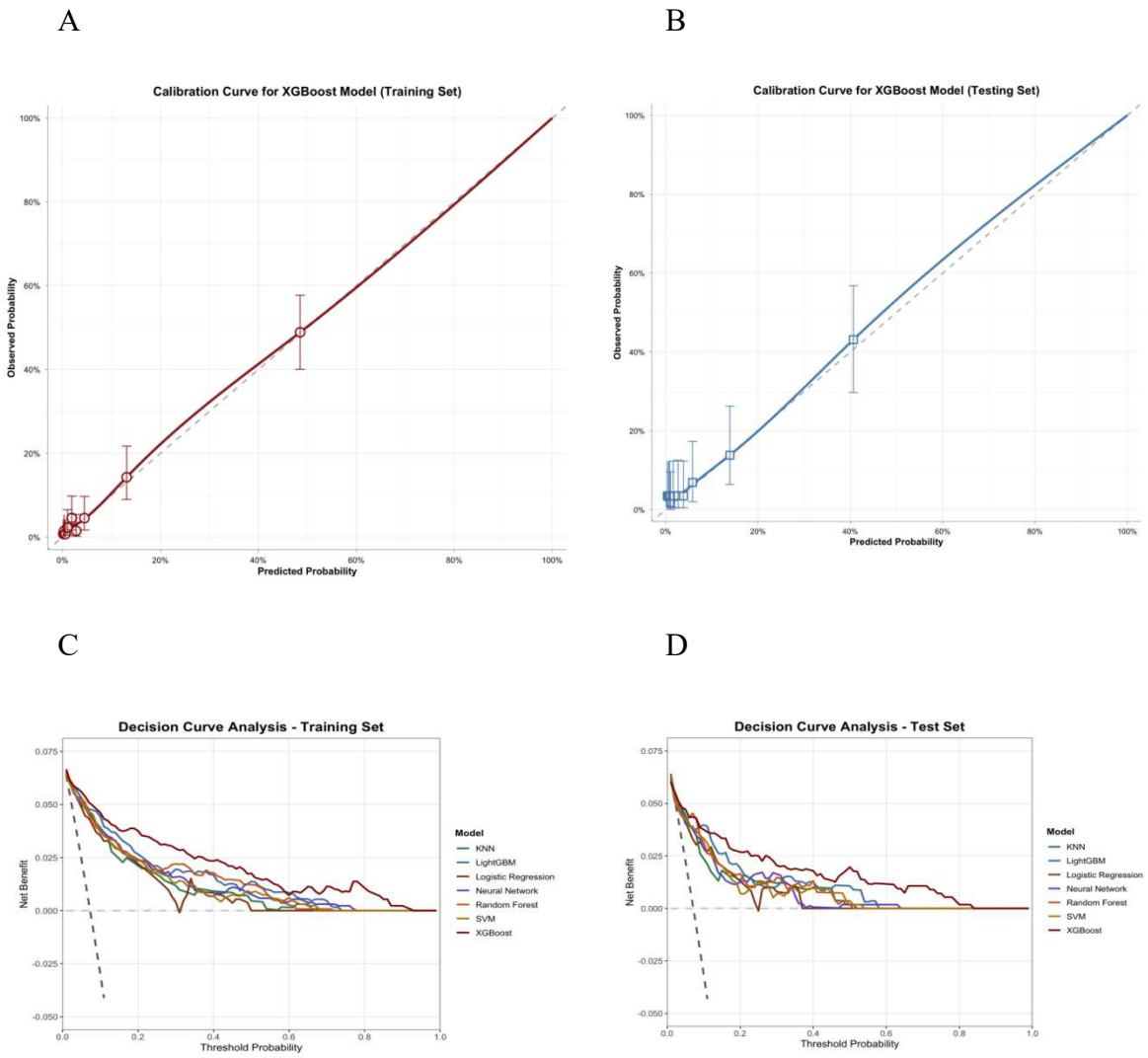


Figure 3. Comparative performance of seven machine learning models and two traditional scales. (A) ROC curves in the training set; (B) ROC curves in the test set; (C) Performance comparison of all models in the training set; (D) Performance comparison of all models in the test set.



534 **Figure 4.** Calibration and clinical utility assessment of the XGBoost model. (A) Calibration
535 curve for the XGBoost model in the training set; (B) Calibration curve for the XGBoost model
536 in the test set; (C) Decision curve analysis in the training set; (D) Decision curve analysis in the
537 test set.

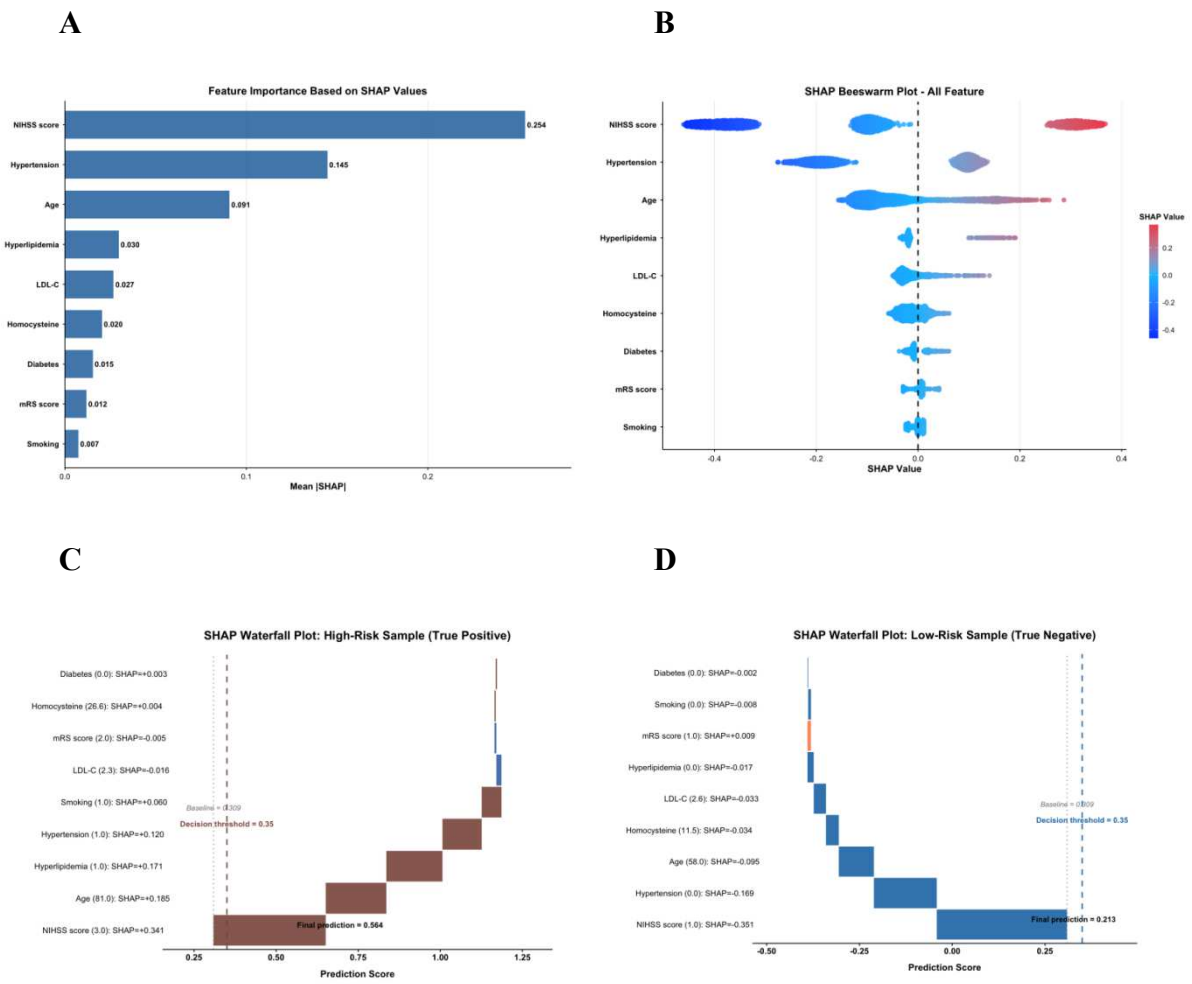


Figure 5. Interpretability of the XGBoost model using SHAP analysis. (A) Feature importance based on mean absolute SHAP values; (B) SHAP summary plot for all core features; (C) SHAP waterfall plot for a high-risk sample (true positive); (D) SHAP waterfall plot for a low-risk sample (true negative).

A B

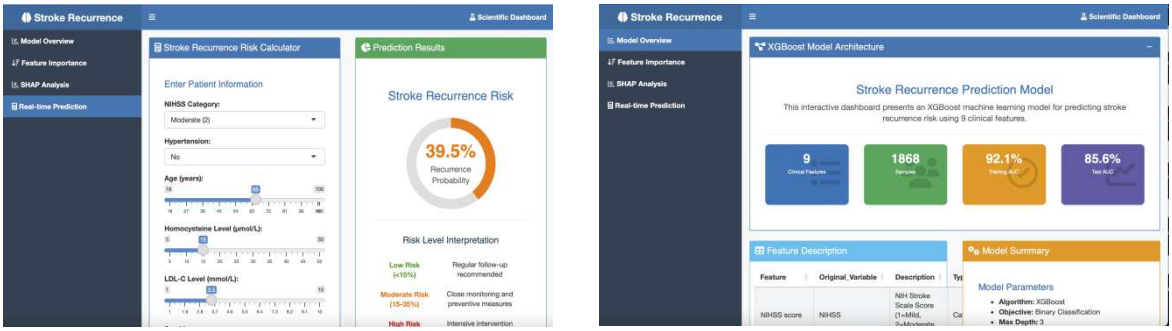


Figure 6. The R Shiny real-time prediction platform for ischemic stroke recurrence. (A) Real-time prediction interface with patient information input and risk assessment results; (B) Model overview panel with XGBoost hyperparameters, dataset characteristics, and core feature descriptions.

546 **Table 1.** The Baseline characteristics of recurrent group and non-recurrent group[M

547 (P25，P75), n (%)]

	Recurrence	Non-recurrence	Statistics	*P*-value
Age (years)	66.2[57.2,75.2]	62.9[55.5,70.4]	3.703	<0.001
Gender				
Female	79(58.0)	1135(65.5)	2.951	0.086
Male	57(42.0)	597(34.5)		
Smoking history	43(31.6)	593(34.2)	5.347	0.021
Drinking history	19(13.9)	322(18.5)	1.804	0.179
Hypertension	81(59.6)	857(49.5)	3.921	0.045
Diabetes	54(39.7)	408(23.5)	3.867	0.047
Hyperlipidemia	30(22.1)	112(6.5)	5.450	0.020
History of carotid	2(1.4)	25(1.4)	0.001	0.980
Family history of	2(1.4)	22(1.3)	0.040	0.842
COPD/Ashma	2(1.4)	5(0.2)	2.944	0.087
Dementia	2(1.4)	13(0.5)	2.980	0.087
Wake-up stroke	8(5.8)	154(8.8)	1.200	0.230
NIHSS (score)			27.430	<0.00
Mild stroke	8(5.9)	701(40.5)		
Moderate stroke	39(28.7)	678(39.2)		
Severe stroke	89(65.4)	353(20.9)		
mRS (score)			10.832	0.001
0-2	77(56.6)	1208(69.75)		
3-5	59(43.4)	524(30.3)		
SBP (mmHg)	148.1[130.0,167.	149.0[132.0,161.	0.669	0.503
DBP (mmHg)	87.8[75.0,97.0]	86.4[78.0,97.0]	1.134	0.257
BMI (Kg/m^2)	23.2[21.3,25.1]	23.5[21.8,25.1]	1.488	0.137
Pulse	78.0[70.0,83.0]	77.2[70.0,84.0]	0.704	0.481
WBC (10^9/L)	7.54[6.11,8.95]	7.30[5.73,8.45]	1.154	0.249
RBC (10^12/L)	4.33[3.91,4.72]	4.53[4.18,4.86]	0.950	0.336
Hb (g/L)	129.0[118.0,141.	136.5[126.0,147.	1.288	0.201
PLT (10^9/L)	230.1[185.0,272.	225.0[179.0,263.	0.836	0.403
Hct (%)	38.4[35.4,42.4]	40.7[38.0,43.9]	1.590	0.111

	Recurrence	Non-recurrence	Statistics	*P*-value
MCV (fL)	89.9[87.8,95.0]	90.6[88.3,94.5]	1.052	0.293
MCH (pg)	29.91[29.00,31.5	30.22[29.43,31.6	1.064	0.287
MCHC (g/L)	329.5[324.0,338.	332.6[325.0,341.	2.631	0.129
Pct (%)	0.24[0.20,0.27]	0.23[0.19,0.27]	0.457	0.648
MPV (fL)	10.6[9.9,11.3]	10.5[9.8,11.2]	0.503	0.615
PLCR (%)	30.2[24.1,34.8]	29.7[23.2,34.7]	0.555	0.579
LYM (10^9/L)	1.64[1.17,1.98]	1.75[1.27,2.08]	1.284	0.199
MONO (10^9/L)	0.61[0.39,0.72]	0.57[0.39,0.61]	0.289	0.773
NEUT (10^9/L)	5.07[3.56,6.30]	4.88[3.50,5.75]	1.004	0.316
TP (g/L)	66.7[62.2,70.3]	67.1[63.5,70.9]	0.783	0.434
ALB (g/L)	38.9[36.2,42.0]	40.0[37.5,42.7]	3.081	0.002
GLB (g/L)	27.8[24.8,30.8]	27.2[24.4,30.0]	1.465	0.143
A/G ratio	1.43[1.25,1.59]	1.50[1.31,1.67]	2.583	0.010
TBIL (μmol/L)	13.24[9.77,15.90	13.81[9.80,16.46	1.126	0.260
DBIL (μmol/L)	3.41[2.23,4.20]	3.57[2.20,4.40]	0.815	0.415
IBIL (μmol/L)	9.81[7.29,12.01]	10.30[7.10,12.22	1.087	0.277
AST (U/L)	25.84[18.00,27.8	24.34[18.00,26.4	0.908	0.364
ALT (U/L)	20.01[12.00,23.6	22.21[13.53,24.7	0.152	0.879
ALP (U/L)	86.4[69.0,97.3]	83.2[66.0,95.6]	1.171	0.242
Urea (mmol/L)	5.21[4.20,6.27]	5.10[4.20,6.60]	0.486	0.627
Cr (mg/dL)	77.81[56.10,86.3	74.67[59.00,83.7	1.205	0.228
UA (μmol/L)	369.5[272.0,412.	368.0[275.9,405.	0.018	0.986
CK (U/L)	113.4[73.1,123.0	117.0[77.0,126.0	0.238	0.812
CK-MB (U/L)	15.31[10.13,19.0	15.22[10.86,17.8	0.070	0.944
LDH (U/L)	191.2[162.2,207.	184.9[157.0,206.	1.540	0.124
α-HBDH (U/L)	146.5[119.7,166.	139.7[118.2,158.	1.987	0.048
TC (mmol/L)	4.49[3.90,4.92]	4.65[4.00,5.24]	1.805	0.071
TG (mmol/L)	1.86[1.13,1.93]	1.77[1.08,2.03]	0.822	0.411
HDL-C (mmol/L)	1.09[0.90,1.23]	1.10[0.92,1.26]	0.626	0.531
LDL-C (mmol/L)	2.73[2.19,3.09]	2.83[2.38,3.37]	2.324	0.020
Non-HDL-C	3.39[2.74,3.86]	3.54[2.93,4.08]	1.768	0.077
AopA (g/L)	1.41[1.02,2.31]	2.32[1.04,2.38]	0.958	0.338
AopB (g/L)	0.88[0.73,1.00]	0.90[0.77,1.06]	1.118	0.264
Lap(a) (mg/L)	123.0[63.0,167.7	139.0[67.1,177.6	0.959	0.338

	Recurrence	Non-recurrence	Statistics	*P*-value
PT activity (%)	100.1[86.8,111.0	103.5[89.8,115.1	1.890	0.056
APTT (s)	25.8[24.0,27.6]	26.2[24.1,28.0]	1.147	0.252
TT (s)	17.2[16.2,18.0]	17.1[16.2,18.1]	0.182	0.856
BG (mmol/L)	6.31[4.83,6.72]	5.80[4.97,7.29]	0.356	0.722
HCY (μmol/L)	14.41[10.51,17.4	12.90[10.31,16.3	1.991	0.047
TOAST			2.271	0.130
CE	36(25.5)	397(22.9)		
LAA	92(67.6)	1124(64.8)		
SAA	6(4.4)	121(7.0)		
SOE/SUE	2(1.5)	90(5.3)		
SPI-II score	10.85[8.32,13.38	7.19[4.89,9.49]	13.131	<0.00
0~3	7(5.1)	109(6.3)		
4~7	44(32.4)	720(41.6)		
8~12	85(62.5)	903(52.1)		
Essen score			3.806	0.051
0~2	89(65.4)	1315(75.9)		
3~5	47(34.6)	417(24.1)		

Notes: NIHSS, National Institutes of Health Stroke Scale (0–4 points for mild stroke; 5–15 points for moderate stroke; ≥16 points for severe stroke); mRS, modified Rankin Scale; COPD, chronic obstructive pulmonary disease; DBP, diastolic blood pressure; BMI, body mass index; WBC, white blood cell count; RBC, red blood cell count; Hb, hemoglobin; PLT, platelet count; Hct, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; Pct, plateletcrit; MPV, mean platelet volume; PLCR, platelet large cell ratio; LYM, lymphocyte count; MONO, monocyte count; NEUT, neutrophil count; TP, total protein; ALB, albumin; GLB, globulin; A/G ratio, albumin/globulin ratio; TBIL, total bilirubin; DBIL, direct bilirubin; IBIL, indirect bilirubin; AST, aspartate aminotransferase; ALT, alanine aminotransferase; ALP, alkaline phosphatase; Cr, creatinine; UA, uric acid; CK, creatine kinase; CK-MB, creatine kinase-MB isoenzyme;

559 LDH, lactate dehydrogenase; α-HBDH, α-hydroxybutyric dehydrogenase; TC, total cholesterol;
560 TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density
561 lipoprotein cholesterol; Non-HDL-C, non-high-density lipoprotein cholesterol; ApoA,
562 apolipoprotein A; ApoB, apolipoprotein B; Lp(a), lipoprotein(a); PT, prothrombin time; APTT,
563 activated partial thromboplastin time; TT, thrombin time; BG, blood glucose;
564 HCY,homocysteine.

565 **Table 2.** Evaluation indicators for the performance of nine recurrence risk predictive

566 models

Model	Accuracy	AUC (95% CI)	Recall	Precision	F1 score	Specificity	Log loss	Brier score	DeLong test
Training set									
LR	0.839	0.868 (0.830-	0.740	0.277	0.403	0.847	0.361	0.105	0.000
RF	0.840	0.882 (0.843-	0.729	0.276	0.400	0.848	0.184	0.051	0.000
SVM	0.830	0.874 (0.837-	0.760	0.268	0.397	0.836	0.190	0.054	0.000
XGBoost	0.879	0.920 (0.892-	0.771	0.352	0.483	0.888	0.152	0.042	——
NN	0.808	0.875 (0.839-	0.802	0.249	0.380	0.809	0.328	0.098	0.000
LightGB	0.810	0.897 (0.870-	0.854	0.259	0.398	0.807	0.244	0.064	0.013
KNN	0.848	0.871 (0.837-	0.635	0.271	0.380	0.865	0.198	0.057	0.002
Essen	0.723	0.558(0.497-	0.550	0.099	0.162	0.753	0.261	0.068	0.000
SPI-II	0.577	0.627(0.572-	0.562	0.096	0.163	0.578	0.256	0.067	0.000
Testing set									
LR	0.825	0.800 (0.721-	0.625	0.231	0.339	0.840	0.376	0.112	0.000
RF	0.823	0.835 (0.763-	0.675	0.239	0.353	0.834	0.205	0.055	0.109
SVM	0.810	0.812 (0.733-	0.725	0.234	0.354	0.817	0.209	0.056	0.000
XGBoost	0.857	0.860 (0.786-	0.725	0.296	0.420	0.867	0.176	0.046	——
NN	0.780	0.802 (0.720-	0.725	0.206	0.320	0.784	0.363	0.113	0.002
LightGB	0.801	0.841 (0.768-	0.800	0.237	0.366	0.802	0.245	0.065	0.849
KNN	0.807	0.820 (0.755-	0.550	0.196	0.289	0.827	0.206	0.058	0.016
Essen	0.742	0.523(0.424-	0.325	0.101	0.154	0.773	0.257	0.066	0.000
SPI-II	0.573	0.525(0.427-	0.575	0.094	0.161	0.573	0.254	0.066	0.000

567

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx