British Journal of Hospital Medicine

Br. J. Hosp. Med. (Lond) **2026**; 87(1): 50394
https://doi.org/10.31083/BJHM50394

Article

Evaluating Predictive Performance of Machine Learning Algorithms That Integrate Routine Clinical Variables With Imaging-Derived Information in Stroke Recurrence Risk

Li Gao[1,2], Shitao Wang[3], Jinlian Li[2], Mingkun Zhang[1,4,*]

[1]Post-doctoral Mobile Research Station, Shandong University of Traditional Chinese Medicine, 250355 Jinan, Shandong, China
[2]Department of Neurology, The Fifth People's Hospital of Jinan, 250022 Jinan, Shandong, China
[3]Department of Interventional Medicine, The Fifth People's Hospital of Jinan, 250022 Jinan, Shandong, China
[4]Department of Neurosurgery, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, 250014 Jinan, Shandong, China
*Correspondence: mingkunzhang3850@126.com (Mingkun Zhang)
Academic Editor: Jessie Welbourne
Submitted: 25 August 2025 Revised: 9 December 2025 Accepted: 18 December 2025 Published: 26 January 2026

Abstract

Aims/Background: Stroke recurrence remains a significant challenge in post-stroke management, with traditional prediction models often showing limited accuracy. This study aims to compare the performance of multiple machine learning (ML) algorithms that integrate routine clinical variables with imaging-derived features in predicting stroke recurrence risk, and to identify the optimal predictive model. **Methods**: This retrospective cohort study enrolled 350 patients with ischemic stroke who were admitted to The Fifth People's Hospital of Jinan between January 2018 and December 2021. Patients were divided into three groups based on the time of first stroke onset: Group A (n = 110), Group B (n = 120), and Group C (n = 120). Routine clinical variables (age, gender, hypertension, and diabetes) and imaging features (infarct size and location) were collected. Four ML-based algorithms—logistic regression, random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost)—were used to construct predictive models. The predictive performance of these models was evaluated by area under the curve (AUC), sensitivity, specificity, and accuracy. **Results**: The XGBoost model showed the superior predictive performance, achieving the highest AUC of 0.86, followed by the random forest model (0.82), support vector machine model (0.78), and logistic regression model (0.75). The most influential predictors for stroke recurrence were found to be infarct size, history of hypertension, and fasting blood glucose levels. **Conclusion**: ML-based algorithms that integrate routine clinical variables with imaging-derived data can predict stroke recurrence risk effectively, with the XGBoost model demonstrating superior predictive performance, which may further support more individualized clinical decision-making.

Keywords: stroke rehabilitation; machine learning; risk assessment; neuroimaging; secondary prevention

1. Introduction

Stroke is a devastating global health burden and remains one of the leading causes of death and long-term disability across all age groups [1,2]. The World Health Organization estimates that over 15 million people experience a stroke each year; approximately 5 million die and 5 million are left with permanent disability [3]. A major challenge in effective stroke management is the significant risk of recurrence. Epidemiological evidence indicates that about 5.7–51.3% of patients experience a second stroke within the first year after the initial event, and the risk can persist for years [4]. Recurrent stroke often results in more severe neurological impairment, increased healthcare costs, and a significant reduction in quality of life for patients and their families [5]. Therefore, early and accurate identification of individuals at high risk of recurrence is not merely a clinical priority but also a critical public health need, enabling individualized secondary prevention strategies to mitigate this risk.

Traditional approaches for predicting the risk of stroke recurrence, such as the Essen Stroke Risk Score (ESRS), the Stroke Prognostic Instrument (SPI), and the ABCD^2 (Age, Blood pressure, Clinical features, Duration of symptoms, Diabetes) score, are widely used in routine clinical care [6]. These models generally rely on a limited set of readily available clinical variables, including age, history of hypertension, diabetes mellitus, atrial fibrillation, and a previous transient ischemic attack (TIA) [7]. While they provide a convenient approach to risk stratification, their predictive performance is often moderate, with validation studies demonstrating area under the curve (AUC) values of 0.6 to 0.7 [8]. This modest accuracy indicates, in part, the limited ability of these strategies to capture the complex, multifactorial biology of stroke, which involves interactions between clinical features, biochemical pathways, and structural brain changes. Moreover, many of these models often do not incorporate detailed neuroimaging information that can provide insights into the severity and anatomical distribution of cerebral damage, all of which are important determinants of recurrence risk.

In recent years, machine learning (ML) has revolutionized various fields of medicine, including diagnostic

imaging, prognostic prediction modeling, and assessment of treatment response [9]. By processing high-dimensional data, identifying non-linear relationships, and extracting complex patterns from large datasets, ML approaches offer a promising alternative to traditional statistical methods for predicting stroke recurrence [10]. Unlike conventional approaches, ML models can integrate diverse data sources, including routine clinical variables, laboratory results, and imaging-derived features, enabling the development of more comprehensive and more accurate prediction tools [11].

Neuroimaging, in particular, holds significant potential for enhancing the prediction of recurrent stroke risk. Computed tomography (CT) and magnetic resonance imaging (MRI) can characterize infarct size and location and detect associated pathologies such as leukoaraiosis, cerebral microbleeds, and carotid artery stenosis [12]. These imaging features can reflect the underlying vascular pathology, the severity of cerebral ischemia, and the burden of silent cerebrovascular disease, all of which are strongly linked to stroke recurrence. For example, larger infarct sizes have consistently been associated with a higher recurrence risk [13], likely indicating more extensive vascular injury and a greater likelihood of unstable atherosclerotic plaques. Similarly, leukoaraiosis, a marker of cerebral small-vessel disease, has been established as an independent predictor of recurrent vascular events [14].

Despite growing interest in applying ML in stroke research, limited studies have performed systematic comparisons of various ML algorithms for predicting stroke recurrence using a combination of routine clinical variables and imaging features. Most published studies have assessed only a single algorithm or have used one data modality alone (e.g., clinical data without imaging, or imaging without detailed clinical data), which limits our understanding of which algorithm and which data integration approach yields the best predictive performance. Additionally, prioritizing and interpreting the most influential predictors of recurrence within an integrated dataset remains crucial, both to enhance model transparency and to generate mechanistic insights that could inform the development of more effective secondary preventive strategies.

Therefore, this study aims to address these gaps by evaluating the performance of four commonly used ML approaches: logistic regression, random forest, support vector machine (SVM), and extreme gradient boosting (XGBoost). Using an integrated dataset that combines routine clinical data with detailed imaging features, the study seeks to determine which algorithm achieves the highest predictive performance for stroke recurrence. Furthermore, the study will identify the most influential predictors of recurrence within the integrated dataset and assess the generalizability of the optimal model across clinically relevant subgroups, such as patients with cortical versus subcortical infarcts. Overall, the findings may support the development of more accurate and clinically useful tools for recurrence risk stratification, enabling more individualized secondary prevention and improved patient outcomes.

2. Methods

2.1 Study Population

This study enrolled 350 patients with ischemic stroke from the Department of Neurology, The Fifth People's Hospital of Jinan, China, between January 2018 and December 2021. Inclusion criteria were as follows: (1) diagnosis consistent with Chinese Stroke Association guidelines for clinical management of ischaemic cerebrovascular diseases: executive summary and 2023 update [15]; (2) first-ever ischemic stroke confirmed by CT or MRI; and (3) availability of complete clinical and imaging data. However, patients were excluded if they had: (1) hemorrhagic stroke; (2) stroke secondary to trauma, tumor, or other non-atherosclerotic causes; (3) severe cognitive impairment or other conditions preventing completion of follow-up.

Patients were categorized into three groups based on the admission period: Group A (January 2018–December 2019), Group B (January 2020–June 2021), and Group C (July 2021–December 2021). This non-uniform time interval design was adopted to account for a hospital-wide transition to a digital medical record system in the later study phase (post–June 2021), which substantially improved the efficiency of patient identification and research recruitment. To ensure balanced sample sizes and baseline characteristics across groups (all $p > 0.05$) while maintaining consistent inclusion criteria, longer intervals were used for Groups A and B (pre-digitalization) to accumulate adequate patients, and a shorter interval was applied for Group C (post-digitalization) to avoid over-recruitment. The primary outcome was stroke recurrence, defined as a new ischemic stroke event confirmed by imaging within one year after the first stroke. A 1-year follow-up was selected because the risk of stroke recurrence is highest during the first year after the initial event, making it a critical window for intensified secondary prevention [16]. The observed difference in monthly enrollment rates across cohorts, including the higher recruitment rate in Group C, likely reflects a hospital-wide transition to a digital medical record system during the later study phase, which significantly improved the efficiency of patient identification and research recruitment while maintaining the same inclusion criteria.

2.2 Data Collection

Two categories of variables, such as routine clinical data and imaging features, were collected for each participant. Routine clinical variables included demographic characteristics (age, gender), comorbidities (hypertension, diabetes, atrial fibrillation, coronary heart disease), laboratory results (fasting blood glucose, total cholesterol, low-density lipoprotein cholesterol, creatinine), and treatment (antiplatelet therapy recorded as a binary variable without

Table 1. Hyperparameter search ranges for machine learning algorithms.

Algorithm	Hyperparameter	Search range	Optimal value (Selected)
Logistic regression (LR)	Penalty (penalty)	1, 2	2
	Regularization strength (C)	0.001, 0.01, 0.1, 1, 10, 100	1
	Solver (solver)	liblinear, saga	liblinear
Random forest (RF)	Number of trees (n_estimators)	50, 100, 200, 500	200
	Max tree depth (max_depth)	3, 5, 10, None	10
	Min samples per leaf (min_samples_leaf)	1, 2, 5	2
SVM	Kernel (kernel)	linear, rbf	rbf
	Penalty parameter (C)	0.1, 1, 10, 100	10
	Gamma (gamma)	scale, auto, 0.01, 0.1, 1	scale
XGBoost	Learning rate (eta)	0.001, 0.01, 0.1, 0.3	0.1
	Max depth (max_depth)	3, 6, 9, 12	6
	Subsample ratio (subsample)	0.6, 0.8, 1.0	0.8
	Min child weight (min_child_weight)	1, 3, 5	3

SVM, support vector machine; XGBoost, extreme gradient boosting; rbf, radial basis function.

specifying the agent or combination regimen, and statin use). Information on formal anticoagulation (e.g., warfarin or direct oral anticoagulants) was not consistently available and was therefore excluded from the analysis. Demographic factors and key comorbidities (hypertension, diabetes, atrial fibrillation, and coronary heart disease) were selected because they are well-established clinical determinants of stroke recurrence.

Imaging features included infarct size (cm^2, measured by CT/MRI), infarct location (cortical, subcortical, or posterior circulation), severity of leukoaraiosis (mild, moderate, severe), and carotid artery stenosis (>50% or not, assessed using ultrasound).

2.3 Machine Learning Models

Four ML algorithms selected for model construction were as follows: (i) Logistic regression (LR), a linear classifier that models the log-odds of binary outcomes, incorporating L1 regularization to reduce overfitting and support feature selection [17]. (ii) Random forest (RF), an ensemble approach that combines multiple decision trees, using bootstrap resampling and random feature selection to enhance robustness and reduce variance [18]. (iii) SVM is a margin-based classifier that identifies an optimal hyperplane to separate classes, using a radial basis function kernel to capture non-linear associations [19]. (iv) XGBoost, a gradient-boosting framework that builds sequential trees with regularization to enhance generalization and minimize prediction error [20].

Feature importance was calculated from each model's internal metric, scoring features based on their average gain across all splits in which they contributed. For benchmarking against traditional risk stratification, the Essen Stroke Risk Score (ESRS) was also calculated for each patient.

2.4 Model Training and Evaluation

The entire cohort was randomly categorized into a training set (70%, n = 245) for model development and an independent testing set (30%, n = 105) for final performance evaluation. All data preprocessing procedures were established using the training data and then applied to the testing data to prevent data leakage. These preprocessing steps included imputation of missing values (median for continuous variables and mode for categorical variables), standardization of continuous variables, one-hot encoding of categorical variables, winsorization of outliers at the 1st and 99th percentiles, and application of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance.

Model hyperparameters were optimized using 5-fold cross-validation within the training set, applying grid search for LR, RF, and SVM, and Bayesian optimization for XGBoost. The hyperparameter search ranges and the optimally selected values are detailed in Table 1. Model performance was determined on the independent testing set using the AUC, sensitivity, specificity, and accuracy. All analyses were performed in Python 3.9 (Python Software Foundation, Beaverton, OR, USA) using scikit-learn (v1.0.2) and XGBoost (v1.5.1) libraries.

2.5 Statistical Analysis

Statistical analyses were conducted using Python (v3.9; Python Software Foundation, Beaverton, OR, USA) with the scikit-learn (v1.0.2) and XGBoost (v1.5.1) libraries, and R (v4.1.2; R Foundation for Statistical Computing, Vienna, Austria) with the tidyverse (v1.3.1) and pROC (v1.18.0) packages. Categorical variables are presented as frequencies and percentages (n, %). Group comparisons were performed using Pearson's chi-square test. Continuous variables: Normality was assessed using the Shapiro-Wilk test, and homogeneity of variances was assessed using Levene's test. Normally distributed contin-

Table 2. Comparison of baseline characteristics of the study participants.

Variable	Group A (n = 110)	Group B (n = 120)	Group C (n = 120)	Test statistic	*p*-value
Age, mean ± SD (years)	65.23 ± 8.34	64.82 ± 7.95	65.51 ± 8.16	F = 0.218	0.805
Male, n (%)	68 (61.82%)	75 (62.50%)	73 (60.83%)	χ^2 = 0.071	0.965
Hypertension, n (%)	82 (74.55%)	89 (74.17%)	91 (75.83%)	χ^2 = 0.097	0.953
Diabetes, n (%)	45 (40.91%)	49 (40.83%)	51 (42.50%)	χ^2 = 0.087	0.958
Atrial fibrillation, n (%)	22 (20.00%)	24 (20.00%)	25 (20.83%)	χ^2 = 0.034	0.983
Coronary artery disease, n (%)	18 (16.36%)	20 (16.67%)	19 (15.83%)	χ^2 = 0.031	0.984
Total cholesterol, mean ± SD (mmol/L)	4.53 ± 0.94	4.62 ± 1.05	4.41 ± 0.86	F = 1.464	0.233
Low-density lipoprotein (LDL) cholesterol, mean ± SD (mmol/L)	2.73 ± 0.74	2.82 ± 0.65	2.61 ± 0.76	F = 2.585	0.077
Creatinine, mean ± SD (µmol/L)	78.23 ± 12.54	77.82 ± 11.95	79.11 ± 13.26	F = 0.328	0.721
Fasting blood glucose, mean ± SD (mmol/L)	5.83 ± 1.24	5.92 ± 1.15	5.71 ± 1.36	F = 0.848	0.429
Antiplatelet therapy, n (%)	95 (86.36%)	103 (85.83%)	104 (86.67%)	χ^2 = 0.036	0.982
Statin use, n (%)	88 (80.00%)	96 (80.00%)	97 (80.83%)	χ^2 = 0.035	0.983
Infarct size, median (IQR) (cm^2)	3.1 (2.0–4.5)	3.0 (1.9–4.3)	3.2 (2.1–4.6)	H = 0.639	0.998
Infarct location, n (%)				χ^2 = 0.108	1.000
Cortical	38 (34.55%)	41 (34.17%)	43 (35.83%)		
Subcortical	60 (54.55%)	66 (55.00%)	65 (54.17%)		
Posterior circulation	12 (10.91%)	13 (10.83%)	12 (10.00%)		
Leukoaraiosis, n (%)				χ^2 = 0.143	0.931
None	42 (38.18%)	46 (38.33%)	45 (37.50%)		
Mild	35 (31.82%)	38 (31.67%)	39 (32.50%)		
Moderate	22 (20.00%)	24 (20.00%)	23 (19.17%)		
Severe	11 (10.00%)	12 (10.00%)	13 (10.83%)		
Carotid stenosis (>50%), n (%)	28 (25.45%)	31 (25.83%)	33 (27.50%)	χ^2 = 0.143	0.931

Continuous variables are presented as mean ± standard deviation (mean ± SD) if normally distributed, or median (interquartile range, IQR) if non-normally distributed (infarct size). For normally distributed variables, group comparisons use one-way ANOVA; for non-normally distributed infarct size, the Kruskal-Wallis H-test is used. Categorical variables are presented as n (%) and compared using Pearson's chi-square test.

uous variables are presented as mean ± standard deviation (mean ± SD) and compared using Student's *t*-test (two groups) or one-way analysis of variance (ANOVA; three or more groups). Non-normally distributed continuous variables are presented as median (interquartile range, IQR) and compared using the Mann-Whitney U test (two groups) or Kruskal-Wallis test (three or more groups). All statistical tests were two-tailed, and a *p*-value of <0.05 was considered statistically significant.

Model calibration, representing agreement between predicted probabilities and observed outcomes, was assessed using the Hosmer-Lemeshow goodness-of-fit test. To further evaluate the key predictors identified by the best-performing model , multivariate logistic regression was performed with adjustment for potential confounders.

3. Results

3.1 Comparison of Baseline Characteristics Across Three Groups

The baseline characteristics of the three groups are summarized in Table 2. No significant differences were found across three groups (Group A, B, and C) regarding age, gender, comorbidities, or imaging features (all $p > 0.05$), indicating that the groups were well balanced at baseline.

3.2 Comparison of Characteristics Between the Training and Testing Sets

Comparison of baseline characteristics between the training set (70% of patients, n = 245) and the testing set (30%, n = 105) is detailed in Table 3. No substantial differences were observed across any variables, including demographic factors, comorbidities, laboratory assessments, treatments, and imaging features (all $p > 0.05$), confirming balanced randomization. This balance ensures the validity of subsequent model training and validation.

3.3 Stroke Recurrence Rate

Stroke recurrence rates across predefined subgroups, including admission-period groups, infarct location, and key clinical risk factors, are summarized in Table 4. Recurrence rates were comparable across the three time-period groups. Conversely, hypertension and carotid stenosis (>50%) were linked to significantly higher recurrence rates, underscoring their role in recurrent stroke risk.

3.4 Predictive Performance of the ML Models

Predictive performance of the four ML models for stroke recurrence is shown in Table 5. Among them, the XGBoost model achieved the highest discrimination, with an AUC of 0.86 (95% confidence interval [CI]: 0.79–0.92),

Table 3. Comparison of baseline characteristics between the training and testing sets.

Variable	Training set (n = 245)	Testing set (n = 105)	Test statistic	*p*-value
Age, mean ± SD (years)	65.14 ± 8.12	65.32 ± 8.05	$t = 0.177$	0.860
Male, n (%)	151 (61.63%)	65 (61.90%)	$\chi^2 = 0.002$	0.962
Hypertension, n (%)	183 (74.69%)	79 (75.24%)	$\chi^2 = 0.012$	0.914
Diabetes, n (%)	101 (41.22%)	44 (41.90%)	$\chi^2 = 0.014$	0.906
Atrial fibrillation, n (%)	49 (20.00%)	22 (20.95%)	$\chi^2 = 0.041$	0.840
Coronary artery disease, n (%)	40 (16.33%)	17 (16.19%)	$\chi^2 = 0.001$	0.975
Total cholesterol, mean ± SD (mmol/L)	4.54 ± 0.92	4.60 ± 0.86	$t = 0.542$	0.588
LDL cholesterol, mean ± SD (mmol/L)	2.72 ± 0.73	2.69 ± 0.65	$t = 0.378$	0.706
Creatinine, mean ± SD (μmol/L)	78.48 ± 12.22	78.03 ± 12.77	$t = 0.302$	0.763
Fasting blood glucose, mean ± SD (mmol/L)	5.85 ± 1.20	5.82 ± 1.18	$t = 0.221$	0.825
Antiplatelet therapy, n (%)	211 (86.53%)	91 (85.71%)	$\chi^2 = 0.020$	0.887
Statin use, n (%)	196 (80.00%)	85 (80.95%)	$\chi^2 = 0.043$	0.836
Infarct size, median (IQR) (cm^2)	3.1 (2.0–4.4)	3.0 (1.9–4.2)	Z = 1258.800	0.715
Infarct location, n (%)			$\chi^2 = 0.010$	0.995
Cortical	85 (34.69%)	37 (35.24%)		
Subcortical	134 (54.69%)	57 (54.29%)		
Posterior Circulation	26 (10.61%)	11 (10.48%)		
Leukoaraiosis, n (%)			$\chi^2 = 0.046$	0.997
None	93 (37.96%)	40 (38.10%)		
Mild	78 (31.84%)	34 (32.38%)		
Moderate	49 (20.00%)	20 (19.05%)		
Severe	25 (10.20%)	11 (10.48%)		
Carotid stenosis (>50%), n (%)	64 (26.12%)	28 (26.67%)	$\chi^2 = 0.012$	0.914

Continuous variables are presented as mean ± standard deviation (mean ± SD) if normally distributed, or median (interquartile range, IQR) if non-normally distributed (infarct size). For normally distributed variables, comparisons use Student's *t*-test; for non-normally distributed infarct size, the Mann-Whitney U-test is used. Categorical variables are presented as n (%) and compared using Pearson's chi-square test.

Table 4. Stroke recurrence rates by time-period groups, infarct location, and clinical characteristics.

Subgroup	Total patients (n)	Recurrent cases (n)	Recurrence rate (%)	Test statistic (χ^2)	*p*-value
Overall cohort	350	78	22.29	—	—
Time-period group				$\chi^2 = 0.020$	0.990
Group A (January 2018–December 2019)	110	24	21.82		
Group B (January 2020–June 2021)	120	27	22.50		
Group C (July 2021–December 2021)	120	27	22.50		
Infarct location				$\chi^2 = 2.104$	0.349
Cortical	122	31	25.41		
Subcortical	191	37	19.37		
Posterior circulation	37	10	27.03		
Comorbidities					
Hypertension	262	69	26.34	$\chi^2 = 9.870$	0.002
No hypertension	88	9	10.23		
Diabetes	145	38	26.21	$\chi^2 = 2.198$	0.138
No diabetes	205	40	19.51		
Atrial fibrillation	71	18	25.35	$\chi^2 = 0.484$	0.487
No atrial fibrillation	279	60	21.51		
Carotid stenosis				$\chi^2 = 6.147$	0.013
>50% Stenosis	92	29	31.52		
≤50% Stenosis	258	49	18.99		

Table 5. Predictive performance of machine learning models in the testing set (n = 105).

Model	AUC (95% CI)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Logistic regression	0.75 (0.67–0.83)	70.5	76.2	74.3
Random forest	0.82 (0.75–0.89)	76.2	80.0	78.1
SVM	0.78 (0.70–0.86)	73.8	78.1	76.2
XGBoost	0.86 (0.79–0.92)	81.0	84.1	83.5
ESRS	0.68 (0.60–0.76)	65.2	70.3	68.9

ESRS, Essen Stroke Risk Score; CI, confidence interval; AUC, area under the curve.

Table 6. Predictive performance in subgroups by infarct location.

Subgroup	Model	AUC (95% CI)
Cortical infarct	XGBoost	0.88 (0.80–0.96)
	Random forest	0.83 (0.74–0.92)
Subcortical infarct	XGBoost	0.84 (0.76–0.92)
	Random forest	0.80 (0.71–0.89)
Posterior circulation	XGBoost	0.81 (0.70–0.92)
	Random forest	0.78 (0.66–0.90)

followed by RF (AUC 0.82, 95% CI: 0.75–0.89), SVM (AUC 0.78, 95% CI: 0.70–0.86), and LR (AUC 0.75, 95% CI: 0.67–0.83). Additionally, the XGBoost model showed the highest sensitivity (81.0%), specificity (84.1%), and overall accuracy (83.5%).

3.5 Calibration of Models

Calibration of all five predictive models, reflecting the agreement between predicted probabilities and observed outcomes, was assessed using the Hosmer-Lemeshow goodness-of-fit test. As shown in **Supplementary Table 1**, all models, including the traditional ESRS, demonstrated good calibration, with non-significant p-values (all $p > 0.05$). These findings indicate close agreement between predicted and observed stroke recurrence risk.

3.6 Subgroup Analysis by Infarct Location

A subgroup analysis stratified by infarct location was conducted to evaluate whether the predictive performance differed across etiologically distinct stroke subtypes, despite comparable overall recurrence rates. To assess the generalizability of the optimal model across these pathophysiologically heterogeneous stroke subtypes, model performance was evaluated individually in subgroups stratified by infarct location: cortical, subcortical, and posterior circulation. As described in Table 6, XGBoost maintained the highest performance across all three subgroups, achieving an AUC of 0.88 (95% CI: 0.80–0.96) for cortical infarcts, 0.84 (95% CI: 0.76–0.92) for subcortical infarcts, and 0.81 (95% CI: 0.70–0.92) for posterior circulation infarcts. Random forest followed as the second-best performer in each subgroup, with AUCs of 0.83, 0.80, and 0.78, respectively.

Table 7. A list of ten most influential predictors of stroke recurrence (XGBoost model).

Predictor	Feature importance
Infarct size	100
History of hypertension	85.2
Fasting blood glucose	78.6
Age	72.1
Carotid artery stenosis (>50%)	68.5
Total cholesterol	62.3
Leukoaraiosis (moderate/severe)	58.9
Diabetes	55.7
Atrial fibrillation	49.2
Antiplatelet therapy	42.8

3.7 Key Predictors of Stroke Recurrence

The ten most influential predictors of stroke recurrence identified by the XGBoost model based on feature importance ranking are listed in Table 7. Infarct size demonstrated the greatest contribution (100.0), followed by a history of hypertension (85.2) and fasting blood glucose (78.6), suggesting crucial roles in recurrence risk prediction.

3.8 Multivariate Logistic Regression for Key Predictors

Multivariate logistic regression findings assessing associations between key predictors and stroke recurrence are shown in Table 8. It revealed that infarct size (odds ratio [OR] = 2.15, 95% CI: 1.52–3.04), hypertension (OR = 1.89, 95% CI: 1.12–3.18), and fasting blood glucose (OR = 1.67, 95% CI: 1.03–2.71) were independently associated with increased recurrence risk of stroke (all $p < 0.05$).

4. Discussion

The present study systematically compared the performance of four machine learning algorithms for predicting stroke recurrence using an integrated set of routine clinical variables and imaging features. Among them, the XGBoost model demonstrated the strongest predictive performance, achieving an AUC of 0.86. The findings underscore the potential of ML-based approaches to enhance risk stratification for stroke recurrence and to address key limitations of traditional prediction models that rely on a narrow set of clinical variables.

Table 8. Multivariate logistic regression of key predictors for stroke recurrence.

Predictor	Regression coefficient	SE	OR	95% CI	*p*-value
Infarct size (per cm^2 increase)	0.77	0.21	2.15	1.52–3.04	<0.001
History of hypertension	0.63	0.26	1.89	1.12–3.18	0.017
Fasting blood glucose (per mmol/L increase)	0.51	0.25	1.67	1.03–2.71	0.038

OR, odds ratio; SE, standard error.

The superior performance of XGBoost compared with logistic regression, random forest, and SVM aligns with previous findings highlighting that gradient-boosting frameworks are well-suited to complex, high-dimensional clinical datasets [21]. A possible explanation for its superior performance is XGBoost's capability to model nonlinear relationships and higher-order interactions among variables, such as the synergistic effect of infarct size and hypertension. For instance, while large infarcts are associated with higher recurrence risk, this effect may be significantly amplified in patients with poorly controlled hypertension, a relationship that linear models such as logistic regression may not capture adequately. This capability is particularly relevant in stroke research, where recurrence risk is determined by a complex interaction of vascular, metabolic, and neuroimaging-related factors.

Integrating imaging-derived features into the predictive models represents a key strength of this study. Traditional models often overlook neuroimaging data because of its analytical complexity and the need for specialized interpretation; however, our results indicate that imaging features, particularly infarct size, contribute significantly to recurrence prediction. Infarct size, ranked as a crucial predictor in the XGBoost model, consistent with previous evidence linking larger infarcts to higher recurrence risk [22]. Larger infarcts usually reflect more severe arterial occlusion, greater ischemic injury, and a higher likelihood of underlying vasculopathy, which all together increase the risk of subsequent cerebrovascular events [23]. Additionally, incorporating markers such as leukoaraiosis and carotid artery stenosis captures the contributions of small-vessel disease and large-artery atherosclerosis, respectively, thereby enhancing the clinical relevance of risk stratification [24].

The identification of hypertension and fasting blood glucose as key predictors reinforces the crucial role of metabolic and vascular risk management in secondary prevention. Hypertension, a well-established driver of stroke pathogenesis, promotes arteriosclerosis, disrupts endothelial function, and increases susceptibility to small vessel occlusion [25]. Similarly, elevated fasting blood glucose levels, even among individuals without a diagnosis of diabetes, may indicate insulin resistance and systemic inflammation, both of which contribute to vascular injury and thrombus formation [26]. Notably, lifestyle-based interventions can significantly improve these metabolic parameters [27]. These findings support current clinical guidelines that emphasize tight blood pressure and glycemic management after stroke, while also highlighting how ML-based models may help identify high-risk individuals who could benefit from more aggressive intervention.

Subgroup analyses revealed that the XGBoost model maintained strong predictive performance across patients with cortical, subcortical, and posterior circulation infarcts, suggesting good generalizability in distinct stroke subtypes with varying etiologies (e.g., large-artery atherosclerosis for cortical, small-vessel disease for subcortical, and vertebrobasilar pathology for posterior circulation). This result is clinically relevant because cortical and subcortical strokes often have distinct etiologies, such as large-artery atherosclerosis and small-vessel disease, and may therefore require tailored preventive strategies [15]. The consistent performance of the model across these subgroups supports its potential ability as a flexible and broadly applicable approach in clinical risk stratification.

Our results also highlight the limitations of traditional risk scores. For example, the ESRS, which relies on variables such as age, hypertension, and diabetes, typically achieves an AUC of about 0.65–0.70 for predicting recurrence [28]. In contrast, the XGBoost model yielded an AUC of 0.86, representing a meaningful improvement in predictive accuracy that could improve identification of high-risk patients. However, ML-based models should be used to complement, not replace, clinical decision-making. While the XGBoost model provides a quantitative risk estimation, clinicians should interpret these findings alongside patient-specific factors, including adherence to medication and lifestyle factors, to guide tailored management.

Several limitations of the study should be considered before interpreting these results. First, the single-center, retrospective design may limit the generalizability of the findings. Variations in clinical practice patterns, imaging acquisition and interpretation, and follow-up procedures across institutions could affect model performance, emphasizing the need for external validation in multicenter cohorts. Second, the study focused on recurrence within the first year of stroke, and longer follow-up is needed to assess how well these models predict late recurrent events. Third, several potentially informative predictors, including genetic markers, lifestyle factors (e.g., smoking status and physical activity), and detailed data on medication adherence, were not included due to unavailability in electronic medical records. Incorporating these variables in future studies may further improve predictive accuracy.

Fourth, while the XGBoost model demonstrated strong performance, the restricted interpretability typical of "black box" models may hinder clinical acceptance without robust explanation frameworks and prospective assessment. Fifth, and importantly, antithrombotic medications were inadequately characterized. The "antiplatelet therapy" was captured only as a binary variable and did not distinguish between single or dual regimens. Crucially, anticoagulant use, which is a critical determinant of recurrence prevention in patients with atrial fibrillation, was not consistently available. The absence of this key confounder likely affected the model's performance and should be addressed in future studies.

Despite these limitations, this study advances our understanding of ML-based stroke recurrence prediction by demonstrating the benefit of integrating routine clinical variables with imaging-derived data. The XGBoost model demonstrated high discriminative performance and consistent outcomes across subgroups, indicating potential application for supporting personalized secondary prevention strategies. However, the single-center, retrospective design and the lack of external validation remain significant limitations and may restrict generalizability. The lack of external validation in diverse, multi-center cohorts represents a significant limitation, potentially affecting the generalizability of our model. Future studies should prioritize external validation to ensure robustness across different patient populations, imaging protocols, and clinical workflows. Furthermore, restricting outcomes to a 1-year recurrence window does not capture late recurrent events, and longer follow-up would strengthen the clinical relevance of the model. Future studies should focus on external validation, incorporating additional predictive variables (such as lifestyle, adherence, and other biologically informative predictors), and develop practical, user-friendly tools to facilitate implementation in routine clinical care.

In summary, machine learning algorithms that integrate routine clinical variables with imaging-derived features can effectively predict stroke recurrence risk, with the XGBoost model offering the highest overall performance. Infarct size, hypertension, and fasting blood glucose were identified as most influential predictors, underscoring the importance of structural neuroimaging and rigorous management of metabolic and vascular risk factors in secondary prevention. These findings support the use of ML-based models as adjuncts to clinical decision-making, with the potential to improve outcomes by facilitating more targeted risk reduction approaches.

5. Conclusion

This study demonstrates that machine learning algorithms integrating routine clinical data and imaging features can predict stroke recurrence risk effectively, with the XGBoost model achieving the highest overall performance. The key predictors, particularly infarct size and a history of hypertension, underscore the significance of structural brain injury and vascular-metabolic dysregulation in driving recurrence risk. Robust performance across cortical, subcortical, and posterior circulation infarct subgroups further supports the model's potential clinical utility in diverse stroke subtypes with distinct pathophysiological mechanisms.

Key Points

- Machine learning models, particularly XGBoost, that integrate both routine clinical and imaging-derived features demonstrate a higher predictive performance for stroke recurrence risk than traditional models.
- Infarct size, a history of hypertension, and fasting blood glucose levels were identified as the most influential predictors of recurrence.
- The XGBoost model maintained robust predictive performance across different stroke subtypes defined by infarct location.
- This study highlights the potential of applying advanced analytical methods and multimodal data for enhancing risk stratification and supporting personalized secondary prevention strategies in stroke survivors.

Availability of Data and Materials

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

LG designed the study. STW and JLL analyzed the data. MKZ performed the study. LG drafted the manuscript. All authors contributed to important editorial changes in the manuscript. All authors read and approved the final version of the manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of The Fifth People's Hospital of Jinan (Approval No. 25-5-16). Informed consent was waived because the study used retrospective, de-identified data from electronic medical records, which involved no intervention or risk to patients. This meets the criteria for waiving informed consent as specified in Article 39 of The Regulations of Ethical Reviews of Biomedical Research Involving Human Subjects of China, which states that retrospective studies using anonymized data with minimal risk to privacy do not require informed consent.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/BJHM50394.

References

[1] Feigin VL, Brainin M, Norrving B, Martins SO, Pandian J, Lindsay P, *et al*. World Stroke Organization: Global Stroke Fact Sheet 2025. International Journal of Stroke. 2025; 20: 132–144. https://doi.org/10.1177/17474930241308142.

[2] Hilkens NA, Casolla B, Leung TW, de Leeuw FE. Stroke. Lancet. 2024; 403: 2820–2836. https://doi.org/10.1016/S0140-6736(24)00642-1.

[3] GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024; 403: 2133–2161. https://doi.org/10.1016/S0140-6736(24)00757-8.

[4] Kolmos M, Christoffersen L, Kruuse C. Recurrent Ischemic Stroke - A Systematic Review and Meta-Analysis. Journal of Stroke and Cerebrovascular Diseases. 2021; 30: 105935. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105935.

[5] Dymm BL, Kwicklis M, Meurer WJ, Shi X, Lisabeth LD. Recurrent stroke arrival time. Journal of Stroke and Cerebrovascular Diseases. 2023; 32: 107069. https://doi.org/10.1016/j.jstrokecerebrovasdis.2023.107069.

[6] Ke L, Zhang H, Long K, Peng Z, Huang Y, Ma X, *et al*. Risk factors and prediction models for recurrent acute ischemic stroke: a retrospective analysis. PeerJ. 2024; 12: e18605. https://doi.org/10.7717/peerj.18605.

[7] Abedi V, Avula V, Chaudhary D, Shahjouei S, Khan A, Griessenauer CJ, *et al*. Prediction of Long-Term Stroke Recurrence Using Machine Learning Models. Journal of Clinical Medicine. 2021; 10: 1286. https://doi.org/10.3390/jcm10061286.

[8] Gladstone DJ, Lindsay MP, Douketis J, Smith EE, Dowlatshahi D, Wein T, *et al*. Canadian Stroke Best Practice Recommendations: Secondary Prevention of Stroke Update 2020. The Canadian Journal of Neurological Sciences. 2022; 49: 315–337. https://doi.org/10.1017/cjn.2021.127.

[9] Karako K, Tang W. Applications of and issues with machine learning in medicine: Bridging the gap with explainable AI. Bioscience Trends. 2025; 18: 497–504. https://doi.org/10.5582/bst.2024.01342.

[10] Colangelo G, Ribo M, Montiel E, Dominguez D, Olivé-Gadea M, Muchada M, *et al*. PRERISK: A Personalized, Artificial Intelligence-Based and Statistically-Based Stroke Recurrence Predictor for Recurrent Stroke. Stroke. 2024; 55: 1200–1209. https://doi.org/10.1161/STROKEAHA.123.043691.

[11] Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischer T. Machine learning and conventional statistics: making sense of the differences. Knee Surgery, Sports Traumatology, Arthroscopy. 2022; 30: 753–757. https://doi.org/10.1007/s00167-022-06896-6.

[12] van Dam-Nolen DHK, Truijman MTB, van der Kolk AG, Liem MI, Schreuder FHBM, Boersma E, et al. Carotid Plaque Characteristics Predict Recurrent Ischemic Stroke and TIA: The PARISK (Plaque At RISK) Study. JACC. Cardiovascular Imaging. 2022; 15: 1715–1726. https://doi.org/10.1016/j.jcmg.2022.04.003.

[13] Arends CM, Liman TG, Strzelecka PM, Kufner A, Löwe P, Huo S, *et al*. Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke. Blood. 2023; 141: 787–799. https://doi.org/10.1182/blood.2022017661.

[14] Moroni F, Ammirati E, Magnoni M, D'Ascenzo F, Anselmino M, Anzalone N, *et al*. Carotid atherosclerosis, silent ischemic brain damage and brain atrophy: A systematic review and meta-analysis. International Journal of Cardiology. 2016; 223: 681–687. https://doi.org/10.1016/j.ijcard.2016.08.234.

[15] Liu L, Li Z, Zhou H, Duan W, Huo X, Xu W, *et al*. Chinese Stroke Association guidelines for clinical management of ischaemic cerebrovascular diseases: executive summary and 2023 update. Stroke and Vascular Neurology. 2023; 8: e3. https://doi.org/10.1136/svn-2023-002998.

[16] Månsson K, Söderholm M, Berhin I, Pessah-Rasmussen H, Ullberg T. The Post-Stroke Checklist: longitudinal use in routine clinical practice during first year after stroke. BMC Cardiovascular Disorders. 2024; 24: 601. https://doi.org/10.1186/s12872-024-04239-6.

[17] Mahmood NH, Kadir DH. Sparsity regularization enhances gene selection and leukemia subtype classification via logistic regression. Leukemia Research. 2025; 150: 107663. https://doi.org/10.1016/j.leukres.2025.107663.

[18] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in Bioinformatics. 2023; 24: bbad002. https://doi.org/10.1093/bib/bbad002.

[19] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20: 273–297. https://doi.org/10.1007/BF00994018.

[20] 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. ACM: New York. 2016. https://doi.org/10.1145/2939672.2939785.

[21] Olaniran OR, Olaniran SF, Allohibi J, Alharbi AA, Alharbi NM. Mixed effect gradient boosting for high-dimensional longitudinal data. Scientific Reports. 2025; 15: 30927. https://doi.org/10.1038/s41598-025-16526-z.

[22] Li G, Feng Z, Zhang H, Zou Y, Xv H, Jiang S. Analysis of influencing factors and interaction effects on stroke recurrence in patients with middle cerebral artery occlusion treated with mechanical thrombectomy. Frontiers in Neurology. 2025; 16: 1580950. https://doi.org/10.3389/fneur.2025.1580950.

[23] Jreij G, Canton G, Hippe DS, Balu N, Yuan C, Cebral J, *et al*. Systematic review of biomechanical forces associated with carotid plaque disruption and stroke. Journal of Vascular Surgery. 2025; 82: 1113–1124.e7. https://doi.org/10.1016/j.jvs.2025.05.014.

[24] Golsari A, Bittersohl D, Cheng B, Griem P, Beck C, Hassenstein A, *et al*. Silent Brain Infarctions and Leukoaraiosis in Patients With Retinal Ischemia: A Prospective Single-Center Observational Study. Stroke. 2017; 48: 1392–1396. https://doi.org/10.1161/STROKEAHA.117.016467.

[25] Zhang C, Li Z, Liu L, Pu Y, Zou X, Yan H, *et al*. The role of hypertension and diabetes mellitus on the etiology of middle cerebral artery disease. Brain and Behavior. 2022; 12: e2521. https://doi.org/10.1002/brb3.2521.

[26] Giacchetti G, Sechi LA, Rilli S, Carey RM. The renin-angiotensin-aldosterone system, glucose metabolism and diabetes. Trends in Endocrinology and Metabolism. 2005; 16: 120–126. https://doi.org/10.1016/j.tem.2005.02.003.

[27] Hörber S, Lehmann R, Fritsche L, Machann J, Birkenfeld AL, Häring HU, *et al*. Lifestyle Intervention Improves Prothrombotic

Coagulation Profile in Individuals at High Risk for Type 2 Diabetes. The Journal of Clinical Endocrinology and Metabolism. 2021; 106: e3198–e3207. https://doi.org/10.1210/clinem/dgab124.

[28] 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. Neuropsychiatric Disease and Treatment. 2023; 19: 2251–2260. https://doi.org/10.2147/NDT.S433383.