


# BMJ Open Development of rapid and effective risk prediction models for stroke in the Chinese population: a cross-sectional study

Yuxin Qiu,<sup>1,2</sup> Shiqi Cheng,<sup>3</sup> Yuhang Wu,<sup>4</sup> Wei Yan,<sup>5</sup> Songbo Hu,<sup>1,2</sup> Yiyin Chen,<sup>5</sup> Yan Xu,<sup>5</sup> Xiaona Chen,<sup>5</sup> Junsai Yang,<sup>1,2</sup> Xiaoyun Chen,<sup>1,2</sup> Huilie Zheng <sup>1,2</sup>

**To cite:** Qiu Y, Cheng S, Wu Y, *et al.* Development of rapid and effective risk prediction models for stroke in the Chinese population: a cross-sectional study. *BMJ Open* 2023;**13**:e068045. doi:10.1136/bmjopen-2022-068045

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-068045>).

Received 09 September 2022  
Accepted 13 February 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

<sup>1</sup>School of Public Health, Nanchang University, Nanchang, Jiangxi, China

<sup>2</sup>Key Laboratory of Preventive Medicine, Nanchang University, Nanchang, Jiangxi, China

<sup>3</sup>Neurosurgery Department, Nanchang University Second Affiliated Hospital, Nanchang, Jiangxi, China

<sup>4</sup>Department of Epidemiology and Health Statistics, Central South University, Changsha, Hunan, China

<sup>5</sup>Institute of Chronic Non-communicable Diseases, Center for Disease Control and Prevention of Jiangxi Province, Nanchang, Jiangxi, China

## Correspondence to

Dr Huilie Zheng;  
[zhenghuilie@ncu.edu.cn](mailto:zhenghuilie@ncu.edu.cn)

## ABSTRACT

**Objectives** The purpose of this study was to use easily obtained and directly observable clinical features to establish predictive models to identify patients at increased risk of stroke.

**Setting and participants** A total of 46 240 valid records were obtained from 8 research centres and 14 communities in Jiangxi province, China, between February and September 2018.

**Primary and secondary outcome measures** The area under the receiver operating characteristic curve (AUC), sensitivity, specificity and accuracy were calculated to test the performance of the five models (logistic regression (LR), random forest (RF), decision tree (DT), extreme gradient boosting (XGBoost) and gradient boosting DT). The calibration curve was used to show calibration performance.

**Results** The results indicated that XGBoost (AUC: 0.924, accuracy: 0.873, sensitivity: 0.776, specificity: 0.916) and RF (AUC: 0.924, accuracy: 0.872, sensitivity: 0.778, specificity: 0.913) demonstrated excellent performance in predicting stroke. Physical inactivity, hypertension, meat-based diet and high salt intake were important prediction features of stroke.

**Conclusion** The five machine learning models all had good predictive and discriminatory performance for stroke. The performance of RF and XGBoost was slightly better than that of LR, which was easier to interpret and less prone to overfitting. This work provides a rapid and accurate tool for stroke risk assessment, which can help to improve the efficiency of stroke screening medical services and the management of high-risk groups.

## INTRODUCTION

Stroke is the leading cause of death and disability worldwide.<sup>1 2</sup> China is one of the countries with the heaviest stroke burden in the world, and the burden of stroke has been increasing in the past 30 years.<sup>3</sup> Over 76% of strokes occur in those without a history of stroke, and mortality and disability associated with strokes significantly affect the lives of patients.<sup>4</sup> The Global Burden of Disease Study reported that stroke incidence decreased by

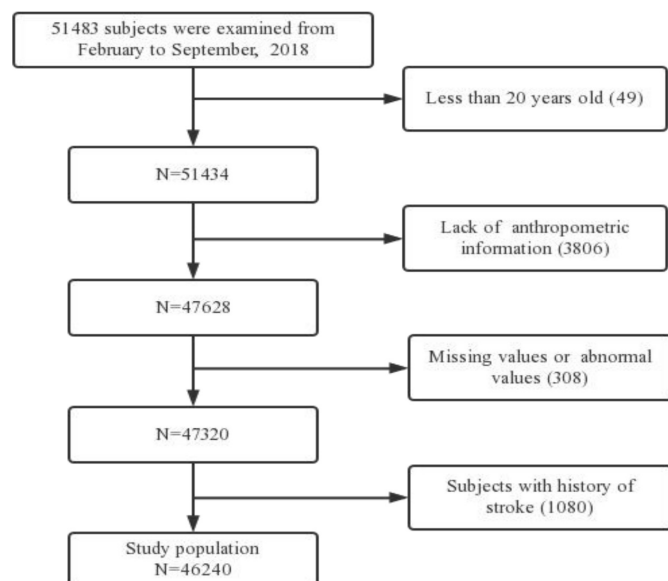
## STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The study used machine learning algorithms with some simple and readily available clinical features for rapid stroke prediction.
- ⇒ The study compared five different algorithms to find the best model, adding to the limited research on stroke risk prediction in China.
- ⇒ Data were collected from 51 483 participants in Jiangxi province using the multistage stratified random cluster sampling method.
- ⇒ The study was cross-sectional, which might have introduced some bias.
- ⇒ Generalisation of study findings to populations of different ages and outside China should be cautious.

12% in countries with practical strategies for preventing cerebrovascular risk factors and good health services in 1990–2010.<sup>5</sup> Prevention of stroke and related risk factors is an essential priority for global public health, especially for low-income and middle-income countries, such as China.

Early stroke screening is an essential means for effective preventive measures. However, the limitations of stroke screening include expensive examination items and an immeasurable workforce. It is unrealistic to ask doctors to make wide-scale diagnoses of stroke using modalities such as ECG, CT and MRI of the brain. In addition, the lack of self-awareness in high-risk individuals makes them want to be tested only when there is a suspected cerebrovascular disease event. To reduce the incidence of stroke, it is vital to develop a simple and accurate method of screening for stroke.

Machine learning has received intense attention for its robust disease prediction capabilities due to its different classification techniques.<sup>6–8</sup> Currently, most machine learning algorithms have been developed as predictive tools for the prognosis of stroke and



**Figure 1** Flow diagram of the study population selected from 8 research centres and 14 counties (cities, districts) in Jiangxi, China.

the occurrence of stroke with other complications, such as using machine learning to predict stroke-associated pneumonia in Chinese patients with acute ischaemic stroke or the outcomes in acute stroke.<sup>9–11</sup> In contrast, there is a lack of research on the construction of stroke risk prediction models, especially in China. Previous Western studies have assessed traditional risk factors (smoking, diabetes, etc). They have developed some risk algorithms to provide valid measures of absolute stroke risk in the general population of patients free of stroke or transient ischaemic attack, as shown by their performance.<sup>12–14</sup> However, it remains questionable whether these models can be reasonably applied to Chinese or other Western populations. A well-known example is the Framingham Stroke Risk Score (FSRS).<sup>15</sup> The FSRS was later modified, particularly for the Chinese population, but the predictive power of the modified model has not been satisfactory.<sup>16</sup>

The development of appropriate disease prediction algorithms is technically challenging. To date, many classical machine learning algorithms have been applied to create a risk assessment for stroke. Li *et al*<sup>17</sup> used the

generalised linear model, Bayes model and decision tree (DT) model to predict the risk of ischaemic stroke and other thromboembolisms in people with atrial fibrillation. Zhang *et al*<sup>18</sup> employed a variety of filter-based feature selection models to improve the ineffective feature selection in existing research on stroke risk detection. Yu *et al*<sup>19</sup> developed a simple, convenient model to predict the risk of stroke among middle-aged and elderly Chinese adults using retrospective cohort datasets. Nevertheless, the sample size is relatively small for developing a prediction model, and the only variables used to build the model are sex, age, hypertension and total cholesterol (TC). Li *et al*<sup>20</sup> developed a logistic regression (LR) model, naïve Bayesian model, Bayesian network model, DT model, neural network model, random forest (RF) model, bagged DT model, voting model and boosting model with DTs to improve stroke risk level classification methods in China. In their study, the outcome of the prediction model was stroke-free individuals at different risk levels determined by the National Stroke Center's screening and intervention project rather than patients who had a stroke. These studies have performed well in stroke prediction, but they cannot fully address the practical issues facing population-level efforts to prevent stroke, especially in China. Therefore, we aim to establish a machine learning-based prediction model to predict stroke occurrence in the population using a sizeable Chinese population and easily obtained and directly observable clinical features.

## MATERIALS AND METHODS

### Study population

This study was supported by the National Stroke Center's screening and intervention project for individuals at high risk of stroke. A total of 51 483 participants (stroke: 18 435; stroke free: 33 048) were recruited in Jiangxi province, China, from February to September 2018. For stroke, we collected electronic health records from eight research centres selected by the National Stroke Center. Stroke was defined by the WHO clinical criteria for stroke.<sup>21</sup> The controls were permanent residents without stroke who had lived in the investigation site for more than 6 months; they were all from the 14 counties (cities,

**Table 1** The choice of hyperparameters for each model

Machine learning models	Hyperparameters	Values to be selected	Optimum value
Random forest	n_estimators (N); max_depth (D); min_samples_leaf (L)	N=1, 2, 3..., 500; D=1, 2, 3..., 30; L=1, 2, 3..., 30	N=291; D=9; L=8
Decision tree	max_depth (D); min_samples_leaf (L); min_samples_split (S)	D=1, 2, 3..., 30; L=1, 2, 3..., 30; S=1, 2, 3..., 30	D=8; L=10; S=4
Gradient boosting decision tree	learning_rate (R); n_estimators (N); max_depth (D)	R=0.01, 0.05, 0.1, 0.2; N=1, 2, 3..., 100; D=1, 2, 3..., 30	R=0.1; N=30; D=6
Extreme gradient boosting	learning_rate (R); n_estimators (N); max_depth (D)	R=0.01, 0.05, 0.1, 0.2; N=1, 2, 3..., 100; D=1, 2, 3..., 30	R=0.1; N=23; D=5

**Table 2** Characteristics of variables in stroke and stroke-free groups

Characteristics	Total (N=46240) No (%)	Stroke-free (N=31880)	Stroke (N=14360)	P value
Sex				<0.001
Men	21 095 (45.6)	12 894 (40.4)	8201 (57.1)	
Women	25 145 (54.4)	18 986 (59.6)	6159 (42.9)	
Age	62.40±11.82	60.64±11.23	66.31±12.17	<0.001
≤55	14 796 (32.0)	11 856 (37.2)	2940 (20.5)	
> 55	31 444 (58.0)	20 024 (62.8)	11 420 (79.5)	
Area				0.495
Urban	23 503 (50.8)	16 238 (50.9)	7265 (50.6)	
Rural	22 737 (49.2)	15 642 (49.1)	7095 (49.4)	
Cardiac causes				<0.001
No	43 335 (93.7)	30 686 (96.3)	12 649 (88.1)	
Yes	2905 (6.3)	1194 (3.7)	1711 (11.9)	
Hypertension				<0.001
No	22 355 (48.3)	19 321 (60.6)	3034 (21.1)	
Yes	23 885 (51.7)	12 559 (39.4)	11 326 (78.9)	
DM				<0.001
No	39 920 (86.3)	28 988 (90.9)	10 932 (76.1)	
Yes	6320 (13.7)	2892 (9.1)	3428 (23.9)	
Smoking				<0.001
No	37 407 (80.9)	26 394 (82.8)	11 013 (76.7)	
Yes	8833 (19.1)	5486 (17.2)	3347 (23.3)	
Alcohol intake				<0.001
No	37 098 (80.2)	25 917 (81.3)	11 181 (77.9)	
Yes	9142 (19.8)	5963 (18.7)	3179 (22.1)	
Physical inactivity				<0.001
No	28 253 (61.1)	25 570 (80.2)	2683 (18.7)	
Yes	17 987 (38.9)	6310 (19.8)	11 677 (81.3)	
High salt intake				<0.001
No	31 717 (68.6)	24 680 (77.4)	7037 (49.0)	
Yes	14 523 (31.4)	7200 (22.6)	7323 (51.0)	
Meat-based diet				<0.001
No	39 920 (86.3)	30 122 (94.5)	9798 (68.2)	
Yes	6320 (13.7)	1758 (5.5)	4562 (31.8)	
Dyslipidaemia				<0.001
No	28 847 (62.4)	21 807 (68.4)	7040 (49.0)	
Yes	17 393 (37.6)	10 073 (31.6)	7320 (51.0)	
High homocysteine				<0.001
No	31 206 (67.5)	22 645 (71.0)	8561 (59.6)	
Yes	15 034 (32.5)	9235 (29.0)	5799 (40.4)	

districts) randomly selected by the multistage cluster sampling method in the catchment areas or nearby areas of the hospitals where cases were recruited. The stroke status was comprehensively judged and ruled out by the neurologist during the interview and investigation after they asked about the history of the stroke, assessed

neurological symptoms and signs, and conducted auxiliary examinations.

We prespecified 13 common independent features related to stroke that have been reported in some previous studies.<sup>22–25</sup> During this process, we fully considered the economy, public acceptance, availability in practice and

**Table 3** Univariate and multivariate logistic regression analysis of variables in predicting stroke

Variables	Univariate analysis		Multivariate analysis	
	OR (95% CI)	P value	OR (95% CI)	P value
Sex (men/women)	0.510 (0.490 to 0.531)	<0.001	0.534 (0.501 to 0.569)	<0.001
Age (>55/≤55)	1.043 (1.041 to 1.045)	<0.001	1.023 (1.020 to 1.025)	<0.001
Cardiac causes (±)	3.476 (3.220 to 3.754)	<0.001	2.140 (1.927 to 2.377)	<0.001
Hypertension (±)	5.743 (5.485 to 6.013)	<0.001	3.906 (3.677 to 4.150)	<0.001
Diabetes mellitus (±)	3.143 (2.977 to 3.318)	<0.001	2.229 (2.068 to 2.403)	<0.001
Smoking (±)	1.462 (1.393 to 1.535)	<0.001	1.110 (1.025 to 1.202)	0.011
Alcohol intake (±)	1.236 (1.177 to 1.297)	<0.001	1.179 (1.095 to 1.267)	<0.001
Physical inactivity (±)	17.636 (16.773 to 18.544)	<0.001	14.338 (13.529 to 15.196)	<0.001
High salt intake (±)	3.567 (3.421 to 3.720)	<0.001	2.188 (2.053 to 2.331)	<0.001
Meat-based diet (±)	7.978 (7.517 to 8.467)	<0.001	4.843 (4.440 to 5.282)	<0.001
Dyslipidaemia (±)	2.251 (2.162 to 2.344)	<0.001	1.947 (1.838 to 2.062)	<0.001
High homocysteine (±)	1.661 (1.594 to 1.731)	<0.001	1.112 (1.047 to 1.181)	<0.001
Area (urban/rural)	0.986 (0.948 to 1.026)	0.495		

whether prevention can be achieved by interfering with these predictive factors. The China Stroke Primary Prevention Trial has shown that a high homocysteine concentration increases the risk of stroke.<sup>26</sup> Eventually, the 13 features included basic predictors (age, sex and area), recognised significant risks (hypertension, smoking, diabetes mellitus, dyslipidaemia and physical inactivity), some modifiable risk factors and characteristics of interest (alcohol intake, high salt intake, meat-based diet, cardiac causes and high homocysteine). Among 51 483 records, a total of 5243 records were excluded. The exclusion criteria were as follows: less than 20 years old; lack of anthropometric information; missing values or abnormal values; and patients with a history of stroke. Finally, a total of 46 240 records were included in this study, as shown in figure 1.

### Data collection

Thirteen variables were included in this study. Cardiac disease was defined as abnormal ECG results or a history of atrial fibrillation, cardiomyopathy, heart failure, ischaemic heart disease, rheumatic heart disease or valvular disease diagnosed by a doctor in secondary or higher hospitals. Hypertension was defined as having a history of being diagnosed with hypertension by a secondary or higher hospital or blood pressure (mean of three measurements) of 140/90 mm Hg or higher. Blood pressure was measured at the time of admission. Diabetes was defined as a history of diabetes or a fasting blood glucose concentration greater than 7.0 mmol/L at the first encounter. Smoking status was defined as cumulative smoking for more than 6 months in a lifetime (current smoking and former smoking). Alcohol intake was classified as never, low or moderate intake and high (more than three times a week and 100 mL each time) intake. Physically active individuals were defined as being involved in moderate or strenuous activity three times or more for 0.5 hours or

more per week or those engaged in moderate or severe physical labour. High salt intake and a meat-biased diet were defined by self-reported daily diet preference for salty taste and appreciation for meat, respectively. For obesity, we assessed body mass index (BMI). Individuals with BMI≥30 were defined as obese.<sup>24</sup> Dyslipidaemia was defined according to the Chinese guidelines for the prevention and treatment of dyslipidaemia in adults as follows<sup>27</sup>: triglycerides≥2.26 mmol/L, TC≥6.22 mmol/L, low-density lipoprotein cholesterol≥4.14 mmol/L and high-density lipoprotein cholesterol<1.04 mmol/L. According to the WHO standard, the average level of homocysteine for healthy adults is 5–15 µmol/L, with a homocysteine level>15 µmol/L representing high homocysteine.<sup>28</sup> The research patients were classified into urban and rural populations based on their areas of residence.

### Patient and public involvement

This research was performed without patient involvement. Patients were not invited to comment on the study design or contribute to the writing or editing of the paper.

### Feature preprocessing

The  $\chi^2$  test and Student's t-test were used for discrete and continuous parameters, respectively. For the independent features of stroke, multivariate LR analysis with backwards stepwise selection was used to calculate the OR with 95% CI. All variables were tested for correlation with each other.

### Construction of machine learning models

In this study, we used five popular machine learning algorithms to predict the probability of a binary outcome (stroke or stroke free): LR,<sup>29</sup> RF,<sup>30</sup> DT,<sup>31</sup> gradient boosting DT<sup>32</sup> and extreme gradient boosting (XGBoost).<sup>33</sup> First, we randomly split our dataset into two groups: the training sets (75%) for machine learning model development and

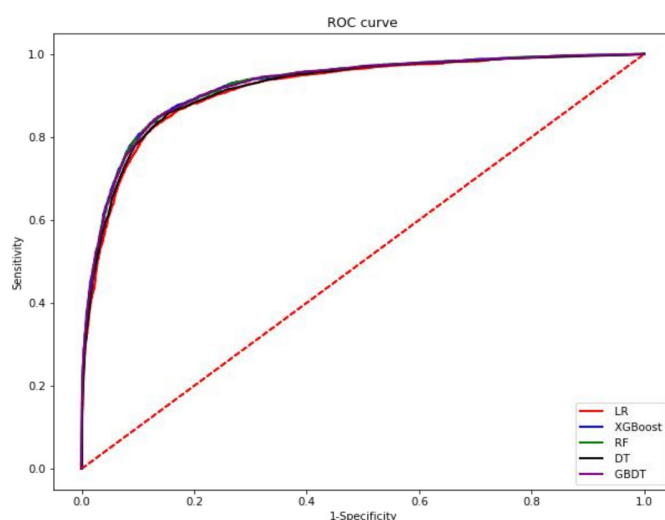


**Table 4** Predictive performance comparison of the five types of machine learning algorithms in the validation sets

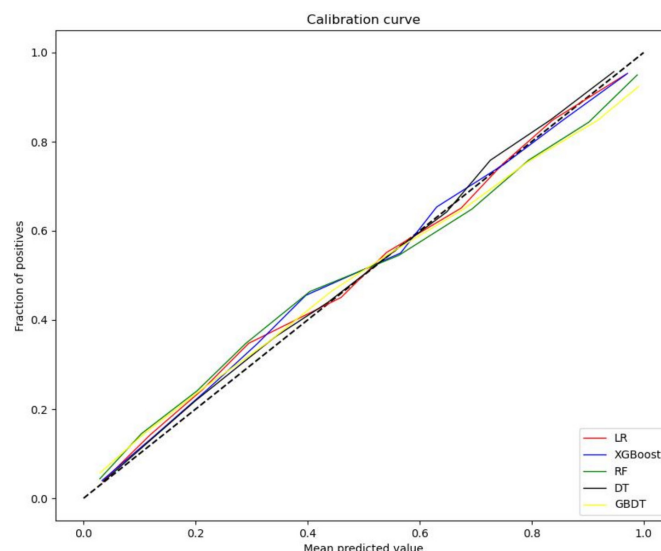
Methods	AUC	Sensitivity	Specificity	Accuracy
LR	0.915	0.762	0.906	0.862
DT	0.903	0.754	0.915	0.866
RF	0.924	0.778	0.913	0.872
XGBoost	0.924	0.776	0.916	0.873
GBDT	0.923	0.765	0.916	0.870

AUC, area under the receiver operating characteristic curve; DT, decision tree; GBDT, gradient boosting decision tree; LR, logistic regression; RF, random forest; XGBoost, extreme gradient boosting.

the validation sets (25%) for performance evaluation. Second, we selected the ranges of hyperparameters to find the best prediction model for each machine learning model. According to the machine learning algorithms, we created a machine learning-based mortality prediction model with hyperparameters for predicting stroke occurrence in the population, which completes the range fitness through grid search using training data. Then, it is evaluated by 10-fold cross-validation. Third, when several hyperparameter combinations were optimal and the choice affected the model's efficiency, we selected the parameter combination that led to the highest efficiency. More details about the features used and their parameter combinations in the models are shown in table 1. Fourth, each machine learning-based model employed the best hyperparameters and was evaluated by the validation sets. The area under the receiver operating characteristic curve (AUC), corresponding sensitivity, specificity and overall accuracy were applied to compare the predictive power of machine learning models; the closer the AUC was to 1, the better the classification model performed.



**Figure 2** Performance characteristic curves for five models (logistic regression (LR), random forest (RF), decision tree (DT), extreme gradient boosting (XGBoost) and gradient boosting decision tree (GBDT)). ROC, receiver operating characteristic.



**Figure 3** Calibration curve showing the agreement between predicted (x-axis) and observed (y-axis) risk of five models. The prediction probability of stroke is divided into 10 bins on average. The diagonal dotted line represents a perfect prediction by an ideal model. DT, decision tree; GBDT, gradient boosting decision tree; LR, logistic regression; RF, random forest; XGBoost, extreme gradient boosting.

The calibration curve was used to show the agreement between the predicted and observed risks of the five models. All variables were tested for correlation with each other, and a heatmap was generated with R (V.4.0.3, R Foundation for Statistical Computing). The R packages 'polycor' and 'ggplot2' were used for correlation analysis; the other statistical analyses were performed with Python (V.3.8, Python Software Foundation). All the results of the models we used in this study could be reproduced by using a fixed random seed.

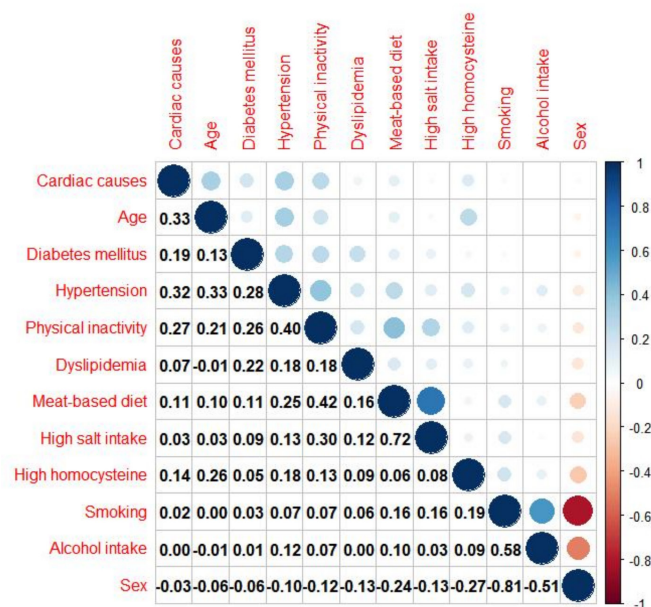
## RESULTS

### Demographic features

A total of 46 240 records (21 095 women and 25 145 men) were selected for this analysis, which included 14 360 records with stroke and 31 880 records without stroke. The average ages were  $66.31 \pm 12.17$  years for patients who had a stroke and  $60.64 \pm 11.23$  years for normal patients. The characteristics of the participants are presented in table 2.

### Univariate and multivariate LR analyses of stroke

In univariable analysis, sex, age, cardiac causes, hypertension, diabetes mellitus, smoking, alcohol intake, physical inactivity, high salt intake, meat-based diet, dyslipidaemia and high homocysteine were all significantly associated with stroke in Jiangxi province ( $p < 0.001$ ). In contrast, there was no significant difference between stroke and stroke-free patients in terms of whether they lived in urban or rural areas. In multivariate LR analysis (table 3), all parameters were included except for area. The results showed that except for women (OR 0.534, 95% CI 0.501



**Figure 4** Results of correlation analysis between all variables.

to 0.569), all the other parameters were independent positive predictors of stroke.

### Performance of machine learning algorithms

Comparisons of the performance of prediction among the five machine learning algorithms models in validation sets are detailed in table 4 and figure 2. The differences between these curves were slight. The performance of XGBoost (AUC: 0.924, accuracy: 0.873, sensitivity: 0.776, specificity: 0.916) and RF (AUC: 0.924, accuracy: 0.872, sensitivity: 0.778, specificity: 0.913) was the best in predicting stroke.

Figure 3 presents a graphical representation of calibration, showing agreement between the predicted and observed risk of the five models. The figure demonstrates that the calibration curves of all models are close to perfect calibration.

All variables were tested for correlation, as shown in figure 4. There was a significant correlation between sex and smoking (correlation coefficient > 0.8).

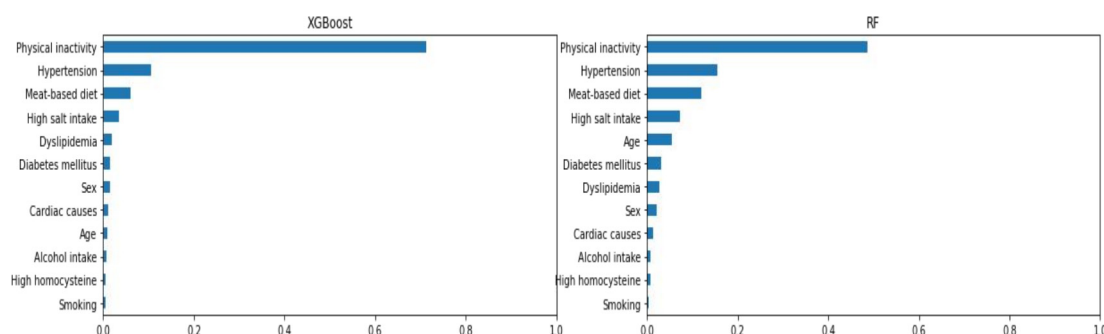
Moreover, according to the information gain values of the five models, the relative importance of variables in

XGBoost and RF is shown in figure 5. We can see there were general evidence trends: physical inactivity contributed the most to stroke, followed by hypertension, a meat-based diet and high salt intake.

### DISCUSSION

In this study, we employed machine learning algorithms to examine the performance of five classifiers and 12 non-invasive and easily obtained clinical features for the rapid and accurate identification of individuals who had a stroke. All models in our study showed very excellent predictive performance, especially RF and XGBoost. This suggests that using machine learning algorithms with some simple and readily available clinical features for rapid stroke prediction is reasonable and feasible. This method is especially suitable for low-income or middle-income areas with heavy stroke burdens, such as China.

RF and XGBoost seem to be the machine learning algorithms of choice in most similar studies.<sup>10 34-36</sup> In the literature, we found that advanced machine learning techniques such as RF and XGBoost modelling can improve the utilisation of information in analytical databases and enable the development and validation of predictive models with better performance.<sup>7</sup> RF and XGBoost showed a considerable degree of predictive power. The RF model was better than XGBoost in accurately detecting patients who had a stroke, whereas the XGBoost model was good at identifying more stroke-free patients. During the training process, the hyperparameters of each algorithm (except for LR) were tuned. We decided not to tune the parameters of the LR model to keep the model specification as simple as possible for comprehensibility. The grid search values were adjusted to optimise the performance of the models. In this study, too many DTs (n=291) in RF required a huge training space and time. In addition, as a black-box model, it cannot control the internal operation of the model for RF, which is not conducive to the interpretation of the model. It was also challenging to avoid complex operating costs for XGBoost. It is worth noting that the classical model, such as LR, also shows solid predictive performance compared with these complex machine learning algorithms. The LR model is easier to use and



**Figure 5** Relative importance ranking of each input variable for prediction of stroke using extreme gradient boosting (XGBoost) and random forest (RF).

**Table 5** Summary of this study and other similar research findings

Reference	Population	Outcome	Variable for prediction	Model	Performance
Our study	Patients who had a stroke from the hospital and stroke free from communities selected by the China national stroke screening and intervention programme	Stroke	Sex, age, cardiac causes, hypertension, diabetes mellitus, smoking, alcohol intake, physical inactivity, high salt intake, meat-based diet, dyslipidaemia, high homocysteine	Logistic regression (LR); random forest (RF); decision tree (DT); gradient boosting decision tree (GBDT); extreme gradient boosting (XGBoost)	AUC: LR=0.905; RF=0.908; DT=0.907; GBDT=0.905; XGBoost=0.907
Yao <i>et al</i> <sup>37</sup>	The China Health and Retirement Longitudinal Survey (CHARLS) for development cohort and the China Health and Nutrition Survey (CHNS) for validation cohort	The 2-year new-set intracranial haemorrhage or ischaemic stroke	Age, hypertension, diabetes, heart disease and smoking	LR	AUC: 0.710 for CHARLS and 0.811 for CHNS
Yu <i>et al</i> <sup>19</sup>	The CHNS	Overall stroke; ischaemic stroke; haemorrhagic stroke	Overall stroke: age, gender, hypertension, total cholesterol; ischaemic stroke: age, gender, hypertension, total cholesterol; haemorrhagic stroke: age, gender, hypertension, body mass index, low-density lipoprotein cholesterol	Cox model	C-index: 0.74 for overall stroke; 0.74 for ischaemic stroke and 0.81 for haemorrhagic stroke
Li <i>et al</i> <sup>20</sup>	Stroke free from communities selected by the China national stroke screening and intervention programme	High-risk level of stroke	Sex, age, drinking history, family history of heart disease, family history of hypertension, family history of diabetes, history of heart disease, heart rhythm and heart murmur	LR; naïve Bayesian (NB); Bayesian network (BN); decision tree (DT); neural network (NN); random forest (RF); bagging model; boosting model; voting model	AUC: LR=99.14%; NB=98.44%; BN=98.41%; DT=99.92%; NN=99.15%; RF=99.94%; Bagging=99.93%; Boosting=99.94%; Voting=99.94%
Lee <i>et al</i> <sup>38</sup>	Farmer cohort from NHIS database in Korea	Stroke	Sex, age, personal history of hypertension, diabetes, current smoking, high $\gamma$ -glutamyl transferase, and metabolic syndrome components (blood pressure, triglycerides, and high-density lipoprotein cholesterol)	Cox model	AUC: 0.760
Lee <i>et al</i> <sup>39</sup>	Participants who had undergone the national health screening obtained from NHIS in Korea	Stroke	Age, body mass index, cholesterol, hypertension, diabetes, smoking status and intensity, physical activity, alcohol drinking, past history (hypertension, coronary heart disease) and family history (stroke, coronary heart disease)	Cox model	AUC: 0.83 for men and 0.82 for women
Chien <i>et al</i> <sup>40</sup>	The Chin-Shan Community Cohort in Taipei, Taiwan, China	Stroke	Age, gender, systolic blood pressure, diastolic blood pressure, family history of stroke, atrial fibrillation and diabetes	Cox model	AUC: 0.772
Dufouil <i>et al</i> <sup>14</sup>	FHS cohort, REGARDS cohort and 3C cohort	Stroke	Age, current smoking, prevalent cardiovascular disease, atrial fibrillation, diabetes mellitus, systolic blood pressure, left ventricular hypertrophy	Cox model	c statistic: 0.74 (FHS), 0.66 (REGARDS), 0.70 (3C) for men and 0.78 (FHS), 0.71 (REGARDS), 0.72 (3C) for women

Continued

Table 5 Continued

Reference	Population	Outcome	Variable for prediction	Model	Performance
Chun <i>et al</i> <sup>41</sup>	The China Kadoorie Biobank individuals	Stroke	lifestyle factors (smoking, alcohol, dietary habits), medical history, physical activity and physical measurements (height, weight, hip and waist circumference, bioimpedance, blood pressure, diabetes and heart rate)	2017 Framingham Stroke Risk Profile (FSRP), a recalibrated and refitted FSRP, Cox, random survival forest (RSF), LR, support vector machine (SVM), gradient boosted tree (GBT) and multilayer perceptron (MLP) models	AUC: FSRP=0.781; refitted RSFP=0.824; Cox=0.829; RSF=0.826; LR=0.831; SVM=0.830; GBT=0.833; MLP=0.831 for men and FSRP=0.772; refitted RSFP=0.825; Cox=0.831; RSF=0.832; LR=0.832; SVM=0.831; GBT=0.836; MLP=0.833 for women

AUC, area under the receiver operating characteristic curve; 3C, 3 Cities; FHS, Framingham Heart Study; NHIS, National Health Insurance Service; REGARDS, Reasons for Geographic and Racial Differences in Stroke.

interpret and less prone to overfitting, but it is sensitive to independent variable multicollinearity. In addition, the correlation analysis results indicated a significant correlation between sex and smoking. However, we still included them in this study because we are more concerned about the predictive power for stroke of the models rather than reporting the impact of stroke.

We have provided more details of similar studies in recent years in table 5. Compared with other studies,<sup>14 19 20 37–41</sup> the models we used have stronger prediction and discrimination performances with higher AUCs, which may be due to the inclusion of more variables in this study. Many machine learning models are sensitive to imbalanced data. The patients we selected included a large number of patients who had a stroke from hospitals, which prevented the classification results from being affected by potential bias. We found that physical inactivity is the most predictive feature of stroke, whether we used the RF or XGBoost models. Physical inactivity is followed by hypertension, meat-based diet and high salt intake. Hypertension has always been considered to be the most important risk factor for stroke,<sup>22 42</sup> which seems to deviate from our results. The results of a large-scale case–control study<sup>23</sup> showed that physical inactivity rather than hypertension was the most important risk factor in China. This also indicates that each region should establish a prediction model with its own geographic and ethnic characteristics based on its own data.<sup>43</sup> In addition, studies<sup>19 40</sup> have reported that age was a significant risk predictor for stroke, whereas it was not highly predictive of stroke in our models. Age group may obscure the contribution of age to stroke in this study. Homocysteine was used as a new predictor to develop a predictive model for stroke. Our results suggest that high homocysteine may not show an important predictive ability for stroke. A meta-analysis reported<sup>44</sup> that elevated homocysteine levels were associated with an increased risk for strokes in different subtypes, which indicated that these stroke risk prediction models built only for overall stroke (ischaemic and haemorrhagic stroke) may underestimate the importance of homocysteine levels for different subtypes of stroke, especially ischaemic stroke. The China Stroke

Primary Prevention Trial has shown that high homocysteine concentration increases the risk of stroke. We included this feature because of interest and ease of detection. Our study showed that high homocysteine was an independent predictor of stroke and that the association with hypertension was not significant, so we retained this feature in the final model. It is a very interesting topic for reflection in future public health work as to whether we will consider a cost-effective or more streamlined version.

The trend in prediction models is to incorporate simplicity and non-invasiveness. In a resource-poor environment, the burden of stroke is disproportionately high.<sup>45 46</sup> The model developed by laboratory testing is difficult to use. Several large-scale studies have demonstrated that far-reaching measures to prevent stroke must involve targeted lifestyle interventions.<sup>22–24 45</sup>

In this study, we used a real dataset of stroke cases from hospitals, and all the cases were diagnosed by doctors, which was more reliable than if the individuals were diagnosed by self-reporting. In addition, data from multiple centres would provide reliable predictive value on how our models identify stroke without selection bias. The models we developed are simple, non-invasive, cost-saving and time-saving, and easy to apply in scenarios other than the clinical setting. We have included enough clinical features to promote stroke screening and prevention in nonprofessional populations. However, some limitations of this study need to be acknowledged. First, this study did not distinguish between ischaemic and haemorrhagic strokes in the diagnosis of stroke. There are some notable differences in risk factors between ischaemic and haemorrhagic stroke.<sup>47</sup> Therefore, more studies with the development of predictive models for ischaemic and haemorrhagic stroke need to be conducted. Second, the models are based on machine learning algorithms, so there may be some difficulties in clinical interpretation of the important features screened out by the models. Third, this is a study based on a province in China, so there may be gaps in population applicability, so it is necessary to include a broader population in future studies. Fourth, the prediction variables obtained retrospectively may leak information to the fitted models, which should be treated



with caution during evaluation. The results should be confirmed in a prospective study. Fifth, the overall accuracy of our model in predicting stroke in the general population is likely to be overly optimistic.

## CONCLUSION

In this study, we demonstrate that the 5 machine learning models developed by using 12 clinical features that are easily obtained and non-invasive all have good predictive and discriminative performance for stroke. The performance of these sophisticated models, such as RF and XGBoost, is slightly better than that of LR, which is easier to interpret and less prone to overfitting. This work provides a rapid and accurate stroke risk assessment tool that can help to improve the efficiency of stroke screening medical services and the management of high-risk populations.

**Acknowledgements** We would like to thank the researchers who participated in this survey.

**Contributors** YQ: Conceptualisation (lead), writing—original draft (lead), formal analysis (lead), writing—review and editing (equal). YW: Writing—original draft (lead), writing—review and editing (equal). SH: Conceptualisation (supporting), formal analysis (supporting), writing—review and editing (equal). WY: Methodology (lead), formal analysis (supporting), writing—review and editing (equal). YC: Conceptualisation (supporting), project administration (equal). YX: Data curation (equal), project administration (equal). XC: Investigation (qual), project administration (equal). JY: Writing—review and editing (equal). XC: Writing—review and editing (equal). SC: Conceptualisation (supporting), supervision (equal). HZ: Conceptualisation (supporting), supervision (equal). YQ is the lead study investigator. HZ is the guarantor.

**Funding** The study was supported by Natural Science Foundation of Jiangxi Province (20202BABL216044), National Natural Science Foundation of China (Grant No.: 81960618), Regional Project of National Natural Science Foundation of China (Grant No.: 82260388), Key projects of Jiangxi Provincial Department of Education (GJJ210118), Project of Jiangxi Provincial Health Commission (202130385) and Key projects of Jiangxi Provincial Administration of Traditional Chinese Medicine (20222017).

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Ethics approval** This study involves human participants and was approved by Xuanwu Hospital Capital Medical University (no. 024 [2015]). Participants gave informed consent to participate in the study before taking part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy concerns.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

## ORCID iD

Huilie Zheng <http://orcid.org/0000-0003-2774-0757>

## REFERENCES

- Campbell BCV, De Silva DA, Macleod MR, et al. Ischaemic stroke. *Nat Rev Dis Primers* 2019;5:70.
- Campbell BCV, Khatir P. Stroke. *Lancet* 2020;396:129–42.
- Wang W, Jiang B, Sun H, et al. Prevalence, incidence, and mortality of stroke in China: results from a nationwide population-based survey of 480 687 adults. *Circulation* 2017;135:759–71.
- Saver JL, Carroll JD, Smalling R, et al. Letter by saver et al regarding article, “ guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline for healthcare professionals from the American heart association/american stroke association.” *Stroke* 2015;46:e85–6.
- Feigin VL, Forouzanfar MH, Krishnamurthi R, et al. Global and regional burden of stroke during 1990–2010: findings from the global burden of disease study 2010. *Lancet* 2014;383:245–54.
- Pei D, Gong Y, Kang H, et al. Accurate and rapid screening model for potential diabetes mellitus. *BMC Med Inform Decis Mak* 2019;19:41.
- Liu W-C, Li Z-Q, Luo Z-W, et al. Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer. *Cancer Med* 2021;10:2802–11.
- Zhu J, Zheng J, Li L, et al. Application of machine learning algorithms to predict central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma. *Front Med (Lausanne)* 2021;8:635771.
- Kostev K, Wu T, Wang Y, et al. Predicting the risk of stroke in patients with late-onset epilepsy: a machine learning approach. *Epilepsy Behav* 2021;122:108211.
- Li X, Wu M, Sun C, et al. Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients. *Eur J Neurol* 2020;27:1656–63.
- Heo J, Yoon JG, Park H, et al. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019;50:1263–5.
- Chambless LE, Heiss G, Shahar E, et al. Prediction of ischemic stroke risk in the atherosclerosis risk in communities study. *Am J Epidemiol* 2004;160:259–69.
- Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of qstroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ* 2013;346:f2573.
- Dufouil C, Beiser A, McLure LA, et al. Revised Framingham stroke risk profile to reflect temporal trends. *Circulation* 2017;135:1145–59.
- D’Agostino RB, Wolf PA, Belanger AJ, et al. Stroke risk profile: adjustment for antihypertensive medication. the framingham study. *Stroke* 1994;25:40–3.
- Huang JY et al. Modified framingham stroke profile in the prediction of the risk of stroke among chinese. *Chinese Journal of Cerebrovascular Diseases* 2013;10:228–32.
- Li X et al. Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. American Medical Informatics Association Annual Symposium (AMIA); 2017
- Zhang Y, Zhou Y, Zhang D, et al. A stroke risk detection: improving hybrid feature selection method. *J Med Internet Res* 2019;21:e12437.
- Yu Q, Wu Y, Jin Q, et al. Development and internal validation of a multivariable prediction model for 6-year risk of stroke: a cohort study in middle-aged and elderly Chinese population. *BMJ Open* 2021;11:e048734.
- Li X, Bian D, Yu J, et al. Using machine learning models to improve stroke risk level classification methods of china national stroke screening. *BMC Med Inform Decis Mak* 2019;19:261.
- Hatano S. Experience from a multicentre stroke register: a preliminary report. *Bull World Health Organ* 1976;54:541–53.
- O’Donnell MJ, Xavier D, Liu L, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* 2010;376:112–23.
- O’Donnell MJ, Chin SL, Rangarajan S, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet* 2016;388:761–75.
- Owolabi MO, Sarfo F, Akinyemi R, et al. Dominant modifiable risk factors for stroke in ghana and nigeria (siren): a case-control study. *Lancet Glob Health* 2018;6:e436–46.
- Cotlarciuc I, Malik R, Holliday EG, et al. Effect of genetic variants associated with plasma homocysteine levels on stroke risk. *Stroke* 2014;45:1920–4.
- Zhao M, Wang X, He M, et al. Homocysteine and stroke risk: modifying effect of methylenetetrahydrofolate reductase C677T polymorphism and folic acid intervention. *Stroke* 2017;48:1183–90.
- Joint Committee for Developing Chinese guidelines on Prevention and Treatment of Dyslipidemia in Adults. Chinese guidelines on prevention and treatment of dyslipidemia in adults. *Zhonghua Xin Xue Guan Bing Za Zhi* 2007;35:390–419.



- 28 Anniwaer J, Liu M-Z, Xue K-D, *et al.* Homocysteine might increase the risk of recurrence in patients presenting with primary cerebral infarction. *Int J Neurosci* 2019;129:654–9.
- 29 Hosmer DW, Lemeshow S. n.d. Applied logistic regression.
- 30 Breiman L. Random forests. *MACH LEARN* 2001;45:5–32.
- 31 Barros RC, Basgalupp MP, de Carvalho ACPLF, *et al.* A hyper-heuristic evolutionary algorithm for automatically designing decision-tree algorithms. GECCO '12; Philadelphia Pennsylvania USA. New York, NY, USA, July 7, 2012
- 32 Cherkassky V, Ma Y. Another look at statistical learning theory and regularization. *Neural Netw* 2009;22:958–69.
- 33 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery, 2016:785–94
- 34 Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 2012;50:1–23.
- 35 Mosca E, Alfieri R, Merelli I, *et al.* A multilevel data integration resource for breast cancer study. *BMC Syst Biol* 2010;4:76.
- 36 Zhang Y, Wang Y, Xu J, *et al.* Comparison of prediction models for acute kidney injury among patients with hepatobiliary malignancies based on xgboost and LASSO-logistic algorithms. *Int J Gen Med* 2021;14:1325–35.
- 37 Yao Q, Zhang J, Yan K, *et al.* Development and validation of a 2-year new-onset stroke risk prediction model for people over age 45 in china. *Medicine (Baltimore)* 2020;99:e22680.
- 38 Lee S, Lee H, Kim HS, *et al.* Incidence, risk factors, and prediction of myocardial infarction and stroke in farmers: a Korean nationwide population-based study. *J Prev Med Public Health* 2020;53:313–22.
- 39 Lee J-W, Lim H-S, Kim D-W, *et al.* The development and implementation of stroke risk prediction model in national health insurance service's personal health record. *Comput Methods Programs Biomed* 2018;153:253–7.
- 40 Chien K-L, Su T-C, Hsu H-C, *et al.* Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. *Stroke* 2010;41:1858–64.
- 41 Chun M, Clarke R, Cairns BJ, *et al.* Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million chinese adults. *J Am Med Inform Assoc* 2021;28:1719–27.
- 42 Lawes CMM, Bennett DA, Feigin VL, *et al.* Blood pressure and stroke: an overview of published reviews. *Stroke* 2004;35:1024.
- 43 Menotti A, Lanti M, Agabiti-Rosei E, *et al.* Riskard 2005. new tools for prediction of cardiovascular disease risk derived from Italian population studies. *Nutr Metab Cardiovasc Dis* 2005;15:426–40.
- 44 He Y, Li Y, Chen Y, *et al.* Homocysteine level and risk of different stroke types: a meta-analysis of prospective observational studies. *Nutr Metab Cardiovasc Dis* 2014;24:1158–65.
- 45 Feigin VL, Roth GA, Naghavi M, *et al.* Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet Neurol* 2016;15:913–24.
- 46 Jia L, Quan M, Fu Y, *et al.* Dementia in china: epidemiology, clinical management, and research advances. *Lancet Neurol* 2020;19:81–92.
- 47 Boehme AK, Esenwa C, Elkind MSV. Stroke risk factors, genetics, and prevention. *Circ Res* 2017;120:472–95.