scientific reports

OPEN

Urban and rural disparities in stroke prediction using machine learning among Chinese older adults

Jingjing Zhu[1], Luotao Lin[2], Lei Si[3], Hailei Zhao[1], Hualing Song[1]✉ & Xianglong Xu[1,4]✉

Stroke is a significant health concern in China. Differences in stroke risk between rural and urban areas have been highlighted in prior research. However, there is a scarcity of studies on urban-rural differences in predicting stroke. This study aimed to develop stroke prediction models, and urban-rural subgroup analyses were conducted to explore disparities in determinants among middle-aged and older adults. We employed nine machine learning algorithms, namely logistic regression (LR), adaptive boosting classifier, support vector machines, extreme gradient boosting, random forest, Gaussian naive Bayes (GNB), gradient boosting machine, light gradient boosting decision machine, and K Nearest Neighbours, using data derived from 9,413 individuals aged 45 years and above obtained from the China Health and Retirement Longitudinal Study (CHARLS) conducted in 2011 to build stroke prediction models and analyze urban-rural subgroups. In the total population, GNB (AUC = 0.76) was the best model for predicting strokes, and the ten most important variables were the time taken for repeated chair stands, the chair height from floor to seat, knee height, creatinine, complete repeated chair stands, mean corpuscular volume, platelet, uric acid, body mass index, and white blood cell. In the rural subgroup, LR and GNB (AUC = 0.76) were the best, and the ten most important variables were the time taken for repeated chair stands, creatinine, platelet, the chair height from floor to seat, knee height, complete repeated chair stands, pulse, white blood cell, maintaining semi - tandem balance statically, and uric acid. In the urban subgroup, LR (AUC = 0.67) was the best, and the ten most important variables were the time taken for repeated chair stands, mean corpuscular volume, maintaining semi - tandem balance statically, uric acid, right-hand grip strength, age, blood urea nitrogen, use of trunk, arms, legs for semi - tandem balance, number of marriages, and night sleep duration. The time taken for repeated chair stands was more critical in the stroke risk model for rural individuals. Uric acid and maintaining semi - tandem balance statically were more critical in the stroke risk model for urban individuals. Our results revealed the importance of knee height and physical function predictors for stroke and highlighted the differences in determinants between urban and rural individuals, proposing targeted stroke prevention and control strategies in different populations in terms of physical function.

Keywords Stroke, Prediction, Machine learning, Urban and rural disparities, Middle-aged and elderly adults

Stroke is a significant health issue in both China and around the world. In 2020, the weighted prevalence and mortality rates of stroke in China were 2.6% and 343.4 per 100,000 person-years, with an estimated 17.8 million prevalent cases and 2.3 million deaths among the middle-aged and elderly population[1]. In recent years, stroke has become the leading cause of death in China, accounting for nearly one-third of global stroke deaths[2]. It is estimated that by 2035, the elderly population aged 60 and above will make up more than 30% of the total population in China, signalling the nation's entry into the era of heavy ageing[3]. Consequently, there will be an increase in both the incidence and mortality of stroke[4], posing challenges for stroke prevention and control. Therefore, it is imperative to take active and vigorous action to mitigate the substantial burden of this disease.

While stroke prevention and control plans and goals are already established in China, attaining these objectives requires targeted interventions and a thorough comprehension of early screening. However, early screening faces many barriers to implementation. From the patients' perspective, unrecognised or ignored stroke symptoms, unconsciousness, hearing problems, and inadequate knowledge hinder accurate stroke assessment[5,6]. Physicians

[1]School of Public Health, Shanghai University of Traditional Chinese Medicine, Shanghai, China. [2]Department of Individual, Family, and Community Education, University of New Mexico, Albuquerque, USA. [3]School of Health Sciences, Western Sydney University, Penrith, Australia. [4]School of Translational Medicine, Monash University, Melbourne, VIC, Australia. ✉email: 99shl@163.com; xianglongxu@shutcm.edu.cn

encounter challenges such as misdiagnosis, inadequate equipment, and personnel shortages[7]. Additionally, language barriers, diverse stroke symptoms, and confusion caused by alcohol or drugs further complicate early stroke screening[8,9]. A study indicates notable disparities in prehospital delays between urban and rural areas, which are associated with longer distances from remote areas to stroke wards, reflecting differences in barriers to screening and diagnosis between urban and rural areas. Moreover, urban-rural disparities affect follow-up treatment access due to factors such as age and socioeconomic status[10]. Therefore, applying predictive models is particularly important to address these obstacles and reduce the impact of urban-rural disparities.

By analysing a large amount of clinical data, predictive models can more accurately predict stroke risk and identify high-risk patients. The rising popularity of machine learning models in stroke prediction is attributed to their capability to tackle complex nonlinear relationships, interactions, and multicollinearity issues that traditional logistic regression cannot[11]. Machine learning algorithms do not require statistical inference or assumptions; they are also self-optimizing and adaptive, making them more accurate and flexible tool for stroke risk prediction. Therefore, their efficiency in stroke prediction is higher[12], allowing physicians and public health workers to create more accurate prevention and control plans earlier[13]. With the continuous development of machine learning technologies, we have more ways and means to construct and optimise predictive models, offering new possibilities to promote early prediction for stroke to reduce the disease burden.

Previous studies have attempted to apply machine learning algorithms to predict stroke risk. A bibliometric analysis showed that most studies have focused on using machine learning to improve stroke risk prediction, diagnosis, and outcome prediction[14]. In studies of stroke risk prediction among the general population, some studies focused on lab variables like blood biomarkers, urine biomarkers and genetic variables[15,16]. Some studies include sociodemographic characteristics, lifestyle factors, diseases, physical examination measurements and blood biomarkers[17–22]. One of the studies lacked physical examination measurements and blood biomarkers[17], and one lacked lifestyle factors[18]. Most studies include a limited number of variables per component, and most physical examination measurements primarily involve body measurements such as BMI, blood pressure, hip circumference, and waist circumference. Only Chang et al. investigated grip strength[22], which may have a significant impact on the risk of cardiovascular disease[23]. A study showed that ideal cardiovascular health is related to better physical function[24]. Notably, there is a lack of research exploring the connection between stroke and other physical functions. In the meantime, stroke risk varies between urban and rural areas. The prevalence of stroke was observed to be marginally higher in urban settings (2.7%) compared to rural areas (2.5%). However, both the incidence rate (485.5 vs. 520.8 per 100,000 person-years) and mortality rate (309.9 vs. 369.7 per 100,000 person-years) were significantly lower in urban locales[1]. Moreover, no studies have conducted subgroup analyses to compare the differences in urban and rural prediction.

Given that current studies have rarely considered the role of physical examination measurements, such as physical function, in stroke and have not examined urban-rural differences in stroke prediction, we used self-reported data, physical examination measurements containing physical function variables, and blood biomarkers to create stroke prediction models. We also conducted urban-rural subgroup analyses to discover urban and rural differences in determinants among middle-aged and older adults.

Results

Characteristics of the study data

Our study included a total of 9,413 participants aged 45 years and above, with 5,033 females and 4,374 males, 7,861 from rural areas and 1,506 from urban areas. In the total population, the median age was 59 years, with an interquartile range (IQR) of 52-66 years. The rural subgroup had a median age of 58 years (IQR 52-65 years), whereas the urban subgroup had a median age of 60 years (IQR 54-67 years). The occurrence of stroke was 2.1% in the total population, 1.9% in rural subgroup, and 3.4% in urban subgroup. Detailed sociodemographic data and other related information can be seen in Table 1 and Supplementary Table 1.

Prediction of stroke

The receiver operating characteristic curves of the nine machine learning algorithms applied to the predictive models were depicted in Fig. 1. GNB demonstrated the highest predictive accuracy (AUC = 0.76), followed by LR (AUC = 0.75), LGBM (AUC = 0.71), XGBoost (AUC = 0.71), RF (AUC = 0.70), GBM (AUC = 0.68), AdaBoost (AUC = 0.64), KNN (AUC = 0.61), and SVM (AUC = 0.51). In subgroup analysis, LR and GNB achieved the highest AUC of 0.76 in the rural subgroup, followed by RF(AUC = 0.67), AdaBoost (AUC = 0.64), XGBoost (AUC = 0.64), GBM (AUC = 0.60), KNN (AUC = 0.60), LGBM (AUC = 0.59) and SVM (AUC = 0.54). LR achieved the highest AUC of 0.67 in the urban subgroup, followed by LGBM (AUC = 0.65), KNN (AUC = 0.63), SVM (AUC = 0.62), GNB (AUC = 0.61), XGBoost (AUC = 0.59), RF (AUC = 0.55), AdaBoost (AUC = 0.54), and GBM (AUC = 0.47).

Important predictors of stroke

The time taken for repeated chair stands, the chair height from floor to seat, knee height, CRE, complete repeated chair stands, MCV, PLT, UA, BMI, and WBC were the ten most important variables in predicting stroke in the total population (Fig. 2a). In the analysis of the rural subgroup, the time taken for repeated chair stands, CRE, PLT, the chair height from floor to seat, knee height, complete repeated chair stands, pulse, WBC, maintaining semi - tandem balance statically, UA were the ten most important variables in predicting stroke (Fig. 2b). In the analysis of the urban subgroup, the time taken for repeated chair stands, MCV, maintaining semi - tandem balance statically, UA, right-hand grip strength, age, BUN, use of trunk, arms, legs for semi - tandem balance, number of marriages, night sleep duration were the ten most important variables in predicting stroke (Fig. 2c). Our study also emphasized the differences in urban and rural subgroups (Fig. 2d). The time taken for repeated chair stands was more critical in the stroke risk model for rural individuals. UA and maintaining semi - tandem

Characteristic	Total population *N* = 9,413	Rural subgroup *N* = 7,861	Urban subgroup *N* = 1,506
Gender			
Female	5,033 (53.5%)	4,287 (54.5%)	723 (48.0%)
Male	4,374 (46.5%)	3,568 (45.4%)	783 (52.0%)
Missing	6 (0.1%)	6 (0.1%)	0 (0.0%)
Age	59 (52,66)	58 (52,65)	60 (54,67)
Education			
Elementary school or lower	6,671 (70.9%)	6,024 (76.6%)	623 (41.4%)
Middle school	1,841 (19.6%)	1,394 (17.7%)	438 (29.1%)
High school or vocational school	779 (8.3%)	425 (5.4%)	341 (22.6%)
College or higher	120 (1.3%)	16 (0.2%)	104 (6.9%)
Missing	2 (0.0%)	2 (0.0%)	0 (0.0%)
Marital status			
Married	8,252 (87.7%)	6,879 (87.5%)	1,328 (88.2%)
Not married, separated, divorced or widowed	1,161 (12.3%)	982 (12.5%)	178 (11.8%)
Number of marriages			
0 time	6 (0.1%)	5 (0.1%)	1 (0.1%)
1 time	8,954 (95.1%)	7,496 (95.4%)	1,415 (94.0%)
2–6 times	384 (4.1%)	298 (3.8%)	83 (5.5%)
Missing	69 (0.7%)	62 (0.8%)	7 (0.5%)
Standard of living			
Poor	1,185 (12.6%)	1,070 (13.6%)	111 (7.4%)
Relatively poor	2,988 (31.7%)	2,445 (31.1%)	529 (35.1%)
Average	4,848 (51.5%)	4,022 (51.2%)	799 (53.1%)
Relatively high	233 (2.5%)	188 (2.4%)	45 (3.0%)
Very high	19 (0.2%)	17 (0.2%)	2 (0.1%)
Missing	140 (1.5%)	119 (1.5%)	20 (1.3%)
Stroke			
No	9,212 (97.9%)	7,711 (98.1%)	1,455 (96.6%)
Yes	201 (2.1%)	150 (1.9%)	51 (3.4%)

Table 1. Characteristics of study participants. Note: Categorical variables are presented as number of participants (%); numeric variables are presented as the median (25%,75%).

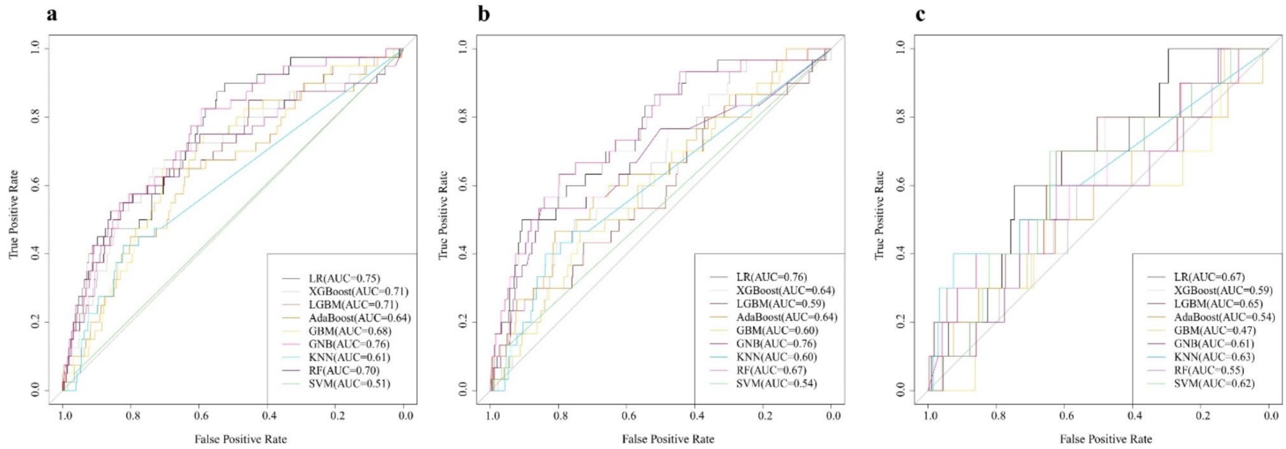

Fig. 1. Receiver operating characteristic curve performance of stroke risk prediction in (**a**) total population, (**b**) rural subgroup, (**c**) urban subgroup. *AUC* area under the curve, *LR* logistic regression, *AdaBoost* adaptive boosting classifier, *SVM* support vector machines, *XGBoost* extreme gradient boosting, *RF* random forest, *GNB* Gaussian naive Bayes, *GBM* gradient boosting machine, *LGBM* light gradient boosting decision machine, *KNN* K Nearest Neighbours.

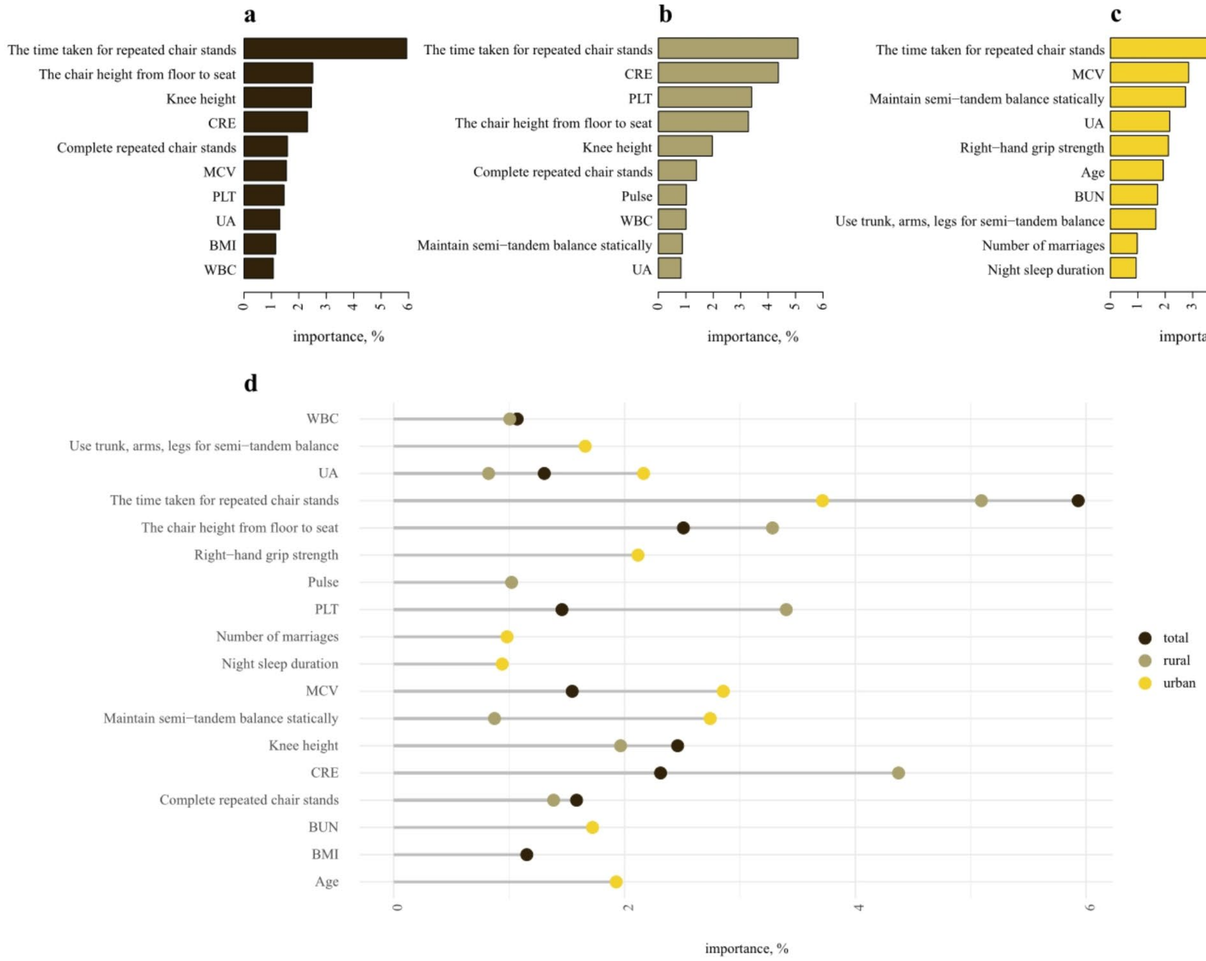

Fig. 2. The top 10 predictors in the prediction of stroke by the best model are (**a**) total population, (**b**) rural subgroup, (**c**) urban subgroup, and (**d**) comparison between groups. *BMI* body mass index, *UA* uric acid, *PLT* platelet, *MCV* mean corpuscular volume, *CRE* creatinine, *BUN* blood urea nitrogen, *WBC* white blood cell.

balance statically were more critical in the stroke risk model for urban individuals. Other variables that did not appear simultaneously in the top ten rankings of the urban-rural subgroup were not compared directly.

Discussion

To our knowledge, this study is the first to utilize machine learning algorithms to develop stroke prediction models among the Chinese urban and rural populations in China. Our findings revealed that machine learning algorithms, which were based on comprehensive data collected from self-reported questionnaires, physical examinations and clinical measurements, performed at an acceptable level to accurately predict stroke in individuals over 45 years old. Our results showed that traditional LR demonstrated superior predictive performance across diverse populations. Our study also demonstrated the importance of physical function predictors collected from physical examination, such as balance abilities and hand grip strength, and the potential predictive value of knee height for stroke, providing new possibilities for prevention and control measures.

Our results showed that machine learning predictive models in the rural subgroup performed better than those in the urban subgroup, reaching acceptable levels. This implies that the machine learning algorithms have higher accuracy in predicting the risk of stroke in the rural subgroup. The variations in lifestyle, health consciousness, and healthcare accessibility between the two populations could account for these results. In China, rural populations aged 45 years and above tend to exhibit a higher frequency of established stroke risk factors compared to urban individuals, such as smoking and excessive alcohol consumption. Additionally, rural areas face challenges in chronic disease management[25]. However, there are no machine learning predictive models of stroke for rural and urban areas. In contrast to other studies, we considered such urban-rural disparities in our study design and analysis, which allowed our models to circumvent obstacles in stroke recognition effectively and facilitated the formulation of more precise and personalized preventive measures tailored to distinct regional populations.

Our findings highlighted the significance of physical function measurements in physical examination, which are related to ideal cardiovascular health[24]. According to the guidelines for the prevention and treatment of stroke in China (2021 Edition)[26], risk factors for stroke are categorized as intervenable and non-intervenable.

Non-interventional factors mainly include age, gender, race, and genetic factors. Intervenable factors include hypertension, glucose metabolism disorders, dyslipidemia, heart disease, asymptomatic carotid atherosclerosis, and lifestyle. Our findings further emphasized the significance of balance abilities and hand grip strength. Concerning balance abilities, we used three tests to analyse the predictive value of balance as a risk factor for stroke: semi-tandem balance, full-tandem balance and repeated chair stands. The last requires lower limb strength and effective balance control[27]. In a study of stroke risk in people with disabilities, those with balance disorders had the highest risk of stroke[28], further supporting the importance of balance in predicting stroke risk. The central nervous system may have an impact on stroke risk, as demonstrated by the ability to maintain balance during daily exercise[28]. Therefore, by assessing and improving balance, we may be able to better predict and reduce the risk of stroke. Most of the studies on balance have been limited to the effect of both on stroke prognosis, and further studies are needed to elucidate their relationship with stroke onset and their underlying mechanisms. What's more, muscle mass has a significant impact on the risk of cardiovascular disease. Research indicates a strong correlation between grip strength and the incidence of coronary heart disease and intracerebral infarction. Furthermore, muscle strength in early adulthood predicts the later risk of heart disease and stroke[23]. These results open new avenues for targeted interventions. Additionally, we identified knee height as a new predictor. The upbringing and living conditions during childhood may affect the risk of certain chronic diseases in later life, and limb length may be an important indicator to observe[29]. Two studies showed that knee height was correlated with diabetes[29,30], and there have been no studies on the relationship between knee height and stroke. The findings complemented those of earlier studies and drew our attention to the potential impact of childhood experiences on stroke.

Furthermore, our study emphasized urban-rural differences in key predictors of stroke. For physical function variables, repeated chair stands test was more critical in the stroke risk model for rural individuals, while balance measurements of semi-tandem were more important in the stroke risk model for urban individuals. We also noted the importance of right-hand grip strength in urban populations. The older adults living in rural areas engage in more physical activities than those living in urban areas[31]. Rural individuals have improved their physical fitness due to long periods of agricultural labour[32], while urban individuals benefit from the convenience of transportation and amenities, which save them from many chores that would otherwise require physical strength. This leads to differences in physical function in elderly adults in the two areas. For some other variables that did not appear in the top ten variable importance in both subgroups, the reasons for the difference in variable importance rankings may be as follows. The differences in upbringing and living conditions between urban and rural areas may affect knee height[29], potentially influencing stroke risk. Societal attitudes contribute to variations in the number of marriages, and disparities in lifestyle and stress levels impact the night sleep duration among individuals in these areas. These findings underscored the need for tailored stroke prevention and intervention strategies based on the specific risk profiles of urban and rural individuals. For rural individuals, promoting physical activities and exercises specifically targeting lower body strength and endurance is important, given the significance of factors related to repeated chair stands. On the other hand, for urban individuals, implementing balance and upper body strength exercises can help improve semi-tandem balance and hand grip strength due to the importance of neuromuscular coordination and muscle strength. Secondly, promoting healthy sleep habits is crucial, as it may be linked to overall health and the risk of stroke.

Limitations

This study has some limitations. First, the study utilised cross-sectional data, which can only estimate the present likelihood of illness, not the future probability of illness. It can also only show the association between the predictors and outcome but not the cause-effect relationship. However, we can argue that the strong association between the identified predictors and outcome could be useful as an indicator in the predictive model. Secondly, although this study combined data from the self-reported questionnaires, physical examinations, and clinical measurements, it did not include all possible factors. Other factors that were also associated with stroke, such as urine markers, could not be included in the model because they were not collected in the CHARLS. In addition, some predictors were self-reported, which can be influenced by participants' recall bias. Thirdly, our study noted new predictive variables, such as upper arm length, but the mechanism of action between them and stroke lacks in-depth confirmation, and future studies should further dissect this relationship. Finally, as the relationships that are valid in the Chinese population may not apply to other populations, the results of our study will require external validation in other populations to guarantee the practicality of the identified predictors.

Methods

Study design and participants

Our research utilised data from the China Health and Retirement Longitudinal Study (CHARLS), a cohort study with national representation from 2011 to 2018. The CHARLS survey used a multistage sampling strategy that covered 28 provinces, 150 counties or districts, and 450 villages or urban communities in China[33]. The study was approved by the Biomedical Ethics Review Committee of Peking University (approval number IRB00001052-11015), and all participants provided written informed consent before being included in the study. All methods were carried out in accordance with relevant guidelines and regulations. Our study used baseline data from 17,705 participants in 2011. We included 9,413 individuals for the study after excluding those who were (1) missing stroke data, (2) missing age data or aged < 45 years, and (3) no physical examination measurements or blood biomarker information.

In addition, participants' family background (urban or rural) was determined by their Hukou status. Hukou is a Chinese household registration system that classifies citizens as rural or urban individuals based on their parents' Hukou registration. It is linked to implementing many social programs and is essential for accessing

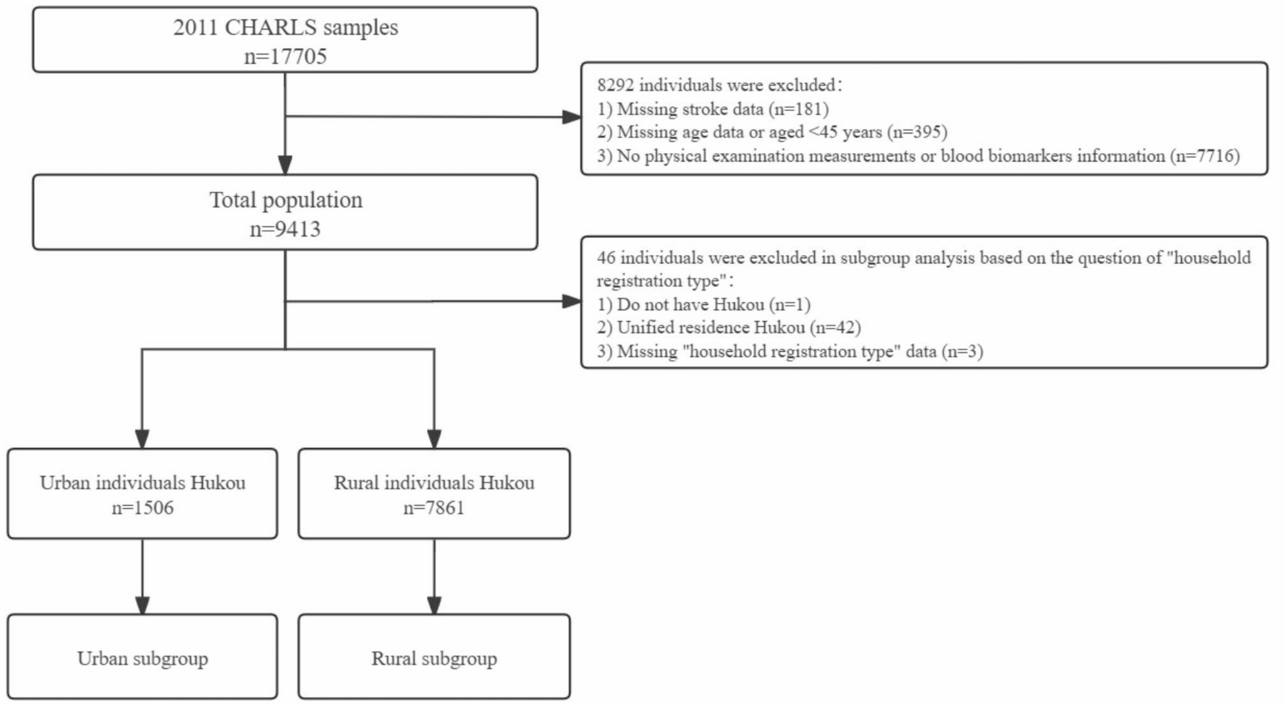
2011 CHARLS samples
n=17705
8292 individuals were excluded:
1) Missing stroke data (n=181)
2) Missing age data or aged <45 years (n=395)
3) No physical examination measurements or blood biomarkers information (n=7716)
Total population
n=9413
46 individuals were excluded in subgroup analysis based on the question of "household registration type":
1) Do not have Hukou (n=1)
2) Unified residence Hukou (n=42)
3) Missing "household registration type" data (n=3)
Urban individuals Hukou
n=1506
Rural individuals Hukou
n=7861
Urban subgroup
Rural subgroup

Fig. 3. Participant encounter inclusion and exclusion diagram.

Categories	Variable types	Variables description
Self-reported data		Gender, age, night sleep duration, hypertension, dyslipidemia, diabetes, struggling with body pains
Physical examination measurements	Body measurements	Upper arm length, knee height, waist circumference, body mass index (BMI)
	Physical function measurements	(1) Balance measurements of semi-tandem [a] maintain semi-tandem balance statically, use trunk, arms, legs for semi-tandem balance (2) Balance measurements of full-tandem [b] maintain full-tandem balance statically, use trunk, arms, legs for full-tandem balance (3) Repeated chair stands test [c] complete repeated chair stands, the time taken for repeated chair stands, the chair height from floor to seat, use trunk arms during repeated chair stands (4) Muscle strength: left-hand grip strength, right-hand grip strength
Blood biomarkers		mean corpuscular volume (MCV), platelet (PLT), blood urea nitrogen (BUN)

Table 2. Selected predictors of stroke. [a]Balance measurements of semi-tandem: stand with the side of the heel of one foot touching the big toe of the other foot for about 10 seconds. [b]Balance measurements of full-tandem: stand with the heel of one foot in front of and touch the toes of the other foot for about [30/60] seconds. 30 seconds for 70 years old or above; 60 seconds for less than 70 years old. [c]Repeated chair stands test: stand up straight and then sit down again at participants' fastest pace five times without stopping in between and without using arms to push off.

government resources[34,35]. In the subgroup analyses, we assigned those with agricultural Hukou to the rural subgroup (n = 7,861) and those with non-agricultural Hukou to the urban subgroup (n = 1,506) based on the question of "household registration type". The remaining individuals who (1) did not have Hukou, (2) had unified residence Hukou, and (3) were missing "household registration type" data were not included in the subgroup analyses. More details on data inclusion and exclusion were shown in Fig. 3.

Predictors and outcome

Our study selected a wide range of predictors, including self-reported data, physical examination measurements, and blood biomarkers (see Table 2). The complete set of predictors of stroke is shown in Supplementary Table 2. The outcome was stroke, which was ascertained through participants' responses to the self-reported question, "Have you been diagnosed with a stroke by a doctor?" Subgroup analyses were conducted based on the criteria, separately for urban and rural subpopulations. Ultimately, we developed stroke prediction models for the total population and subgroups using three categories of predictors.

Analysis

Data processing

We used R 4.1.3 for data processing. The variables with less than 30% missing values were imputed through the mice package with the random forest method. The dataset was partitioned into training and testing sets using an 80%-20% split. To ensure consistent random results, random seeds were utilized during the data splitting process, and the data order was shuffled to minimize sample correlation. The dataset presented a notable disproportion in the number of samples between negative and positive stroke outcomes, leading to a potential issue of data imbalance. Such imbalance could bias the model towards predicting the majority class, possibly causing overfitting or compromising predictive accuracy. We resampled the imbalanced dataset to address this issue. Specifically, we applied oversampling techniques to the training dataset, aiming to achieve a more equitable representation of both negative and positive outcomes.

Machine learning algorithms

We utilised nine common machine learning algorithms, namely logistic regression (LR), adaptive boosting classifier (AdaBoost), support vector machines (SVM), extreme gradient boosting (XGBoost), random forest (RF), gaussian naive Bayes (GNB), gradient boosting machine (GBM), light gradient boosting decision machine (LGBM), and K Nearest Neighbours (KNN) to construct risk prediction models for stroke. Machine learning algorithms were conducted with Python 3.8.12. LR, AdaBoost, SVM, RF, GNB, GBM, LGBM and KNN were built using the scikit-learn library in Python. XGBoost was built using the XGBoost library in Python. To identify the best hyperparameters for machine learning algorithms, we employed a five-fold cross-validation approach along with Bayesian optimization.

Performance measurement

The performance of the model was evaluated using the area under the curve (AUC) metrics, which represent the area under the receiver operating characteristic (ROC) curve. A higher AUC value indicates a better prediction effect of model[36]. An AUC value ranging from 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is regarded as excellent, and any value exceeding 0.9 is categorized as outstanding[11].

Statistical descriptive analysis

The median (M) and interquartile range (IQR) were used for numerical data to characterize the distribution. For categorical data, it was described using frequency and percentage. All statistical analyses were performed in R 4.1.3.

Data availability

The datasets generated during and/or analysed during the current study are available in the CHARLS repository, https://charls.pku.edu.cn/.

Received: 13 May 2024; Accepted: 18 February 2025
Published online: 25 February 2025

References

1. Tu, W. J. et al. Estimated burden of stroke in China in 2020. *JAMA Netw. Open*. **6**, e231455. https://doi.org/10.1001/jamanetworkopen.2023.1455 (2023).
2. Wang, W. et al. Prevalence, incidence, and mortality of stroke in China: results from a nationwide Population-Based survey of 480 687 adults. *Circulation* **135**, 759–771. https://doi.org/10.1161/CIRCULATIONAHA.116.025250 (2017).
3. Division, A. Transcript of the September 20, 2022 press conference by the NHSRC. http://www.nhc.gov.cn/xcs/s3574/202209/ee4dc20368b440a49d270a228f5b0ac1.shtml (2022).
4. Wu, S. et al. Stroke in China: advances and challenges in epidemiology, prevention, and management. *Lancet Neurol.* **18**, 394–405. https://doi.org/10.1016/S1474-4422(18)30500-3 (2019).
5. Bakke, I., Lund, C. G., Carlsson, M., Salvesen, R. & Normann, B. Barriers to and facilitators for making emergency calls - a qualitative interview study of stroke patients and witnesses. *J. Stroke Cerebrovasc. Dis.* **31**, 106734. https://doi.org/10.1016/j.jstrokecerebrovasdis.2022.106734 (2022).
6. Mackay, E., Theron, E. & Stassen, W. The barriers and facilitators to the telephonic application of the FAST assessment for stroke in a private emergency dispatch centre in South Africa. *Afr. J. Emerg. Med.* **11**, 15–19. https://doi.org/10.1016/j.afjem.2020.11.002 (2021).
7. Meng, Z. et al. Development and validation of a LASSO prediction model for better identification of ischemic stroke: A Case-Control study in China. *Front. Aging Neurosci.* **13**, 630437. https://doi.org/10.3389/fnagi.2021.630437 (2021).
8. Zhao, J. & Liu, R. Stroke 1-2-0: a rapid response programme for stroke in China. *Lancet Neurol.* **16**, 27–28. https://doi.org/10.1016/S1474-4422(16)30283-6 (2017).
9. Hodell, E. et al. Paramedic perspectives on barriers to prehospital acute stroke recognition. *Prehosp Emerg. Care*. **20**, 415–424. https://doi.org/10.3109/10903127.2015.1115933 (2016).
10. Buus, S. et al. Urban-rural inequalities in IV thrombolysis for acute ischemic stroke: A nationwide study. *Eur. Stroke J.* **23969873241244591** https://doi.org/10.1177/23969873241244591 (2024).
11. Xu, X. et al. Web-Based Risk Prediction Tool for an Individual's Risk of HIV and Sexually Transmitted Infections Using Machine Learning Algorithms: Development and External Validation Study. *J. Med. Internet. Res.* **24**. https://doi.org/10.2196/37850 (2022).
12. Chahine, Y. et al. Machine learning and the conundrum of stroke risk prediction. *Arrhythmia Electrophysiol. Rev.* **12**. https://doi.org/10.15420/aer.2022.34 (2023).
13. Xi, Y., Wang, H. & Sun, N. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. *Front. Cardiovasc. Med.* **9**, 1025705. https://doi.org/10.3389/fcvm.2022.1025705 (2022).
14. Pahwa, B., Tayal, A. & Garg, K. Contributions of machine learning in the management of stroke: A bibliometric analysis of the 50 most cited articles. *World Neurosurg.* **184**, 152–160. https://doi.org/10.1016/j.wneu.2024.01.059 (2024).
15. Alanazi, E. M., Abdou, A. & Luo, J. Predicting risk of stroke from lab tests using machine learning algorithms: development and evaluation of prediction models. *JMIR Form. Res.* **5**, e23440. https://doi.org/10.2196/23440 (2021).
16. Theofilatos, K., Korfiati, A., Mavroudi, S., Cowperthwaite, M. C. & Shpak, M. Discovery of stroke-related blood biomarkers from gene expression network models. *BMC Med. Genomics*. **12**, 118. https://doi.org/10.1186/s12920-019-0566-8 (2019).
17. Qiu, Y. et al. Development of rapid and effective risk prediction models for stroke in the Chinese population: a cross-sectional study. *BMJ Open*. **13**, e068045. https://doi.org/10.1136/bmjopen-2022-068045 (2023).

18. Lolak, S., Attia, J., McKay, G. J. & Thakkinstian, A. Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: retrospective cohort study. *JMIR Cardio*. **7**, e47736. https://doi.org/10.2196/47736 (2023).
19. Dritsas, E. & Trigka, M. Stroke risk prediction with machine learning techniques. *Sens. (Basel)*. **22** https://doi.org/10.3390/s22134670 (2022).
20. Hong, C. et al. Predictive accuracy of stroke risk prediction models across black and white race, sex, and age groups. *Jama* **329**, 306–317. https://doi.org/10.1001/jama.2022.24683 (2023).
21. Chun, M. et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *J. Am. Med. Inform. Assoc.* **28**, 1719–1727. https://doi.org/10.1093/jamia/ocab068 (2021).
22. Chang, H. W. et al. Ischemic stroke prediction using machine learning in elderly Chinese population: the Rugao longitudinal ageing study. *Brain Behav.* **13**, e3307. https://doi.org/10.1002/brb3.3307 (2023).
23. Silventoinen, K., Magnusson, P. K., Tynelius, P., Batty, G. D. & Rasmussen, F. Association of body size and muscle strength with incidence of coronary heart disease and cerebrovascular diseases: a population-based cohort study of one million Swedish men. *Int. J. Epidemiol.* **38**, 110–118. https://doi.org/10.1093/ije/dyn231 (2009).
24. Fan, D. et al. Cardiovascular health profiles, systemic inflammation, and physical function in older adults: A population-based study. *Arch. Gerontol. Geriatr.* **109**, 104963. https://doi.org/10.1016/j.archger.2023.104963 (2023).
25. Prevention, T. N. B. o. D. C. a. Report on the Nutrition and Chronic Diseases Status of Chinese Residents 2020: BeijingThe peoples medical publishing house,. (2021).
26. China, N. H. C. o. t. P. s. R. o. Guidelines for the prevention and treatment of stroke in China (2021). Edition http://www.nhc.gov.cn/yzygj/s3593/202108/50c4071a86df4bfd9666e9ac2aaac605/files/674273fa2ec049cc97ff89102c472155.pdf (2021).
27. Medina-Mirapeix, F., Crisostomo, M. J., Martín San Agustín, R. & Sánchez-Martínez, M. P. Prognostic value of balance performance for improvements of community ambulation among stroke patients: a cohort study. *Eur. J. Phys. Rehabil Med.* **58**, 171–178. https://doi.org/10.23736/s1973-9087.21.06996-3 (2022).
28. Inchai, P., Tsai, W. C., Chiu, L. T. & Kung, P. T. Incidence, risk, and associated risk factors of stroke among people with different disability types and severities: A National population-based cohort study in Taiwan. *Disabil. Health J.* **14**, 101165. https://doi.org/10.1016/j.dhjo.2021.101165 (2021).
29. He, B. et al. Upper arm length and knee height are associated with diabetes in the middle-aged and elderly: evidence from the China health and retirement longitudinal study. *Public. Health Nutr.* **26**, 190–198. https://doi.org/10.1017/s1368980022001215 (2023).
30. Palloni, A., McEniry, M., Wong, R. & Peláez, M. The tide to come: elderly health in Latin America and the Caribbean. *J. Aging Health*. **18**, 180–206. https://doi.org/10.1177/0898264305285664 (2006).
31. Zhu, W., Chi, A. & Sun, Y. Physical activity among older Chinese adults living in urban and rural areas: A review. *J. Sport Health Sci.* **5**, 281–286. https://doi.org/10.1016/j.jshs.2016.07.004 (2016).
32. Chen, X., Lin, Z., Gao, R., Yang, Y. & Li, L. Prevalence and associated factors of falls among older adults between urban and rural areas of Shantou City, China. *Int. J. Environ. Res. Public. Health*. **18**. https://doi.org/10.3390/ijerph18137050 (2021).
33. Zhao, Y., Hu, Y., Smith, J. P., Strauss, J. & Yang, G. Cohort profile: the China health and retirement longitudinal study (CHARLS). *Int. J. Epidemiol.* **43**, 61–68 (2014).
34. Wei, J. M., Li, S., Claytor, L., Partridge, J. & Goates, S. Prevalence and predictors of malnutrition in elderly Chinese adults: results from the China health and retirement longitudinal study. *Public. Health Nutr.* **21**, 3129–3134. https://doi.org/10.1017/s1368980018002227 (2018).
35. Wang, G. et al. Determinants of COVID-19 vaccination status and hesitancy among older adults in China. *Nat. Med.* **29**, 623–631. https://doi.org/10.1038/s41591-023-02241-7 (2023).
36. Du, Z. et al. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and Machine-Learning methods: model development and performance evaluation. *JMIR Med. Inf.* **8**, e17257. https://doi.org/10.2196/17257 (2020).

Acknowledgements

This work was supported by Artificial Intelligence-Driven Reform of Scientific Research Paradigms to Empower Disciplinary Advancement Program, Shanghai University of Traditional Chinese Medicine in 2021 (2021LK008), and Shanghai University of Traditional Chinese Medicine in 2024 (KECJ2024019). The funders were not involved in the study design, data collection, analysis, decision to publish, or manuscript preparation. The authors extend their thanks to the CHARLS team and all participants for their valuable time and effort in enabling this study.

Author contributions

ZJ and XX conceived and designed the study. XX and ZJ established the models and coding. ZJ, XX contributed to data cleaning. ZJ wrote the first draft and edited the manuscript. LL, SL, SH and ZH contributed to the manuscript revision. All authors contributed to the preparation of the manuscript and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91157-y.

Correspondence and requests for materials should be addressed to H.S. or X.X.

Reprints and permissions information is available at www.nature.com/reprints.

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