


Designing machine learning for big data: A study to identify factors that increase the risk of ischemic stroke and prognosis in hypertensive patients

DIGITAL HEALTH
Volume 10: 1–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241288833
journals.sagepub.com/home/dhj



Lingmin Gong^{1, #}, Shiyu Chen^{1, #}, Yuhui Yang^{1, #}, Weiwei Hu¹, Jiaxin Cai¹, Sitong Liu¹, Yaling Zhao¹, Leilei Pei¹, Jiaojiao Ma³ and Fangyao Chen^{1, 2} 

Abstract

Background: Ischemic stroke (IS) accounts large amount of stroke incidence. The aim of this study was to discover the risk and prognostic factors that affecting the occurrence of IS in hypertensive patients.

Method: Study data were obtained from the Medical Information Mart for Intensive Care (MIMIC)-IV database. To avoid biased factors selection process, several approaches were studied including logistic regression, elastic net regression, random forest, correlation analysis, and multifactor logistic regression methods. And seven different machine-learning methods are used to construct predictive models. The performance of the developed models was evaluated using AUC (Area Under the Curve), prediction accuracy, precision, recall, F1 score, PPV (Positive Predictive Value) and NPV (Negative Predictive Value). Interaction analysis was conducted to explore potential relationships between influential factors.

Results: The study included 92,514 hypertensive patients, of which 1746 hypertensive patients experienced IS. The Gradient Boosted Decision Tree (GBDT) model outperformed the other prediction model terms of prediction accuracy and AUC values in both ischemic and prognosis cases. By using the SHapley Additive exPlanations (SHAP), we found that a range of factors and corresponding interactions between factors are important risk factors for IS and its prognosis in hypertensive patients.

Conclusion: The study identified factors that increase the risk of IS and poor prognosis in hypertensive patients, which may provide guidance for clinical diagnosis and treatment.

Keywords

IS, risk factor, prognosis, machine learning, explainable learning

Submission date: 25 May 2024; Acceptance date: 17 September 2024

Introduction

The global disease burden of stroke over the past decades has not been favorable. Stroke was reported to be the second leading cause of death, and the third leading cause of combined death and disability.¹ Published systematic analysis revealed that among the 10 diseases with the greatest number of neurological disability-adjusted life years (DALYs) in 2021, stroke accounted for the largest share globally and in 19 of the 21 GBD regions.² The global burden of stroke increased substantially from 1990 to 2019 by 70.0% in stroke events and 143.0% in DALYs.³ Previously published study showed that nearly 15% to

¹Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi, China

²Department of Radiology, The First Affiliate Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

³Department of Neurology, Xi'an Gaoxin Hospital, Xi'an, Shaanxi, China

[#]These authors contribute equally to this study.

Corresponding author:

Fangyao Chen, Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi 710061, China.
Email: chenfy@xjtu.edu.cn

Jiaojiao Ma, Department of Neurology, Xi'an Gaoxin Hospital, Xi'an, Shaanxi 710075, China.
Email: jma9211@126.com



30% of stroke survivors would experience lifelong disability, while 20% require at least three months of hospital care after stroke.⁴ Stroke imposes significant costs of care and costs associated with lost productivity on patients.⁵

Ischemic stroke (IS) is a general term for necrosis of brain tissue due to narrowing or occlusion of the arteries supplying blood to the brain (carotid and vertebral arteries) and insufficient blood supply to the brain.⁶ The vast majority of stroke incidence are IS.⁵ The prevalence of IS in China has also increased significantly,⁷ data from the Hospital Quality Inspection System (HQMS) in 2019 showed that 1672 tertiary hospitals admitted IS accounted for 82.6% of strokes.⁸

Hypertension, one of the major comorbidities of stroke, is prevalent in the stroke population and is the most important modifiable risk factor for stroke.⁹ Therefore, large numbers of research on the factors influencing the onset and prognosis of IS in hypertensive patients. Currently, studies on the occurrence and prognosis of IS are often based on hospital follow-up data, and public databases. Among those databases, the Medical Information Mart for Intensive Care (MIMIC) database has adequate number of patients with IS, with comprehensive and complete records of the various indexes, which is eligible for the study purposes.

There are a number of machine-learning prediction studies on stroke in which a number of meaningful influences have been identified, as well as effective machine-learning models. These provide medicine with a basis for stroke prevention and treatment.^{10,11} The current application of the MIMIC database related to IS are mainly conducted for two types. For the first type, the analysis of all-cause mortality and risk of death in IS patients using traditional statistical models.^{12,13} These studies focused on a single factor and only on patients admitted to the intensive care unit (ICU). However, a large proportion of IS patients are not admitted to the ICU on their initial admission. If we lose this part of the data, meaningful information may be lost and the research results may be biased. Another type of studies has been conducted to develop predictive models.^{14,15} However, such studies often suffer from incomprehensive study factors, non-robustness in factor screening and machine-learning-based modeling, as well as the lack of interpretability.¹⁶

Meanwhile, relevant experimental studies have found that age is a risk factor for IS in hypertensive patients, but the effect of age was not the same in male and female studies.¹⁷ It was also found that changes in BMI were associated with high prevalence of IS and diabetes mellitus, while diabetic patients had an increased risk of IS.¹⁸ This type of literature has inspired us to explore the interactions between the factors that influence the development of IS in hypertensive patients, and exploring the interactions between these factors can help to provide a more comprehensive understanding of the pathogenesis of IS, leading

to the development of more effective preventive and therapeutic strategies. However, few data studies have explored the interactions between variables.

Previous studies have focused on healthy and diabetic populations, and few have looked at stroke risk in people with hypertension.¹⁹ Therefore, in this study, we included as many potential factors related to risk and prognosis of IS in hypertensive patients as possible, used a more robust factor screening strategy. And to address the potential heterogeneity, an interpretable machine-learning approach was used to explore the effects of interactions on outcomes, in order to better support the clinical diagnosis and treatment.

Methods

Sources of data

Research data was obtained from the MIMIC-IV database (version 2.0, <https://physionet.org/content/mimiciv/2.0/>). The database is a high-quality, publicly available dataset consisting of clinical information on patients who were at Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019.²⁰

Study population

Patients diagnosed with hypertension and IS were included according to ICD9 and ICD10, specific codes selected are available in Supplementary Materials. The exclusion criteria were as follows: (a) age <18 years; (b) absence of hypertension at the time of first admission; (c) IS occurring before hypertension; and (d) death occurring before IS.

Data collection and study outcome

This study was a retrospective observational study based on public database conducted at the Xi'an Jiaotong University Department of Epidemiology and Biostatistics from January 2023 to April 2024.

The extraction in the MIMIC-IV database was carried out in the Navicate Premium (version 16.0.11, Premium Soft, Hong Kong, China) software platform and Structured Query Language (SQL) were applied to extract the data.

The list of extracted factors is also provided in Supplementary Materials. The endpoint for studies is the occurrence of IS for hypertensive patients and prognosis for hypertensive patients with IS.

Data pre-processing

The study data extracted from databases contain missing values. Factors with $\geq 50\%$ amusingness in the extracted data were not included in subsequent analyses. Other

missing data were filled in using multiple imputation method. We also transformed the continuous factor BMI in the database into ordinal factor. The use of three medications, simvastatin, rosuvastatin, and atorvastatin, were combined into statins use, which was recorded as 1 whenever one of these medications was used.

When different features have different orders of magnitude, larger values may dominate the learning process of the model, resulting in a model that is less sensitive to features with smaller values. Standardization solves this problem by keeping all features within a similar range. So we use this method for deep neural network (DNN) model that rely on gradient descent.²¹ Decision trees (DT) and their derived models usually do not require standardization or normalization of the data as they are based on comparisons of eigenvalues rather than numerical magnitudes.

Factor selection

In order to capture a broader range of factors relevant to the outcome and ensure that the factors included are well-represented and generalizable, we used three different methods for the factor screening in the training set²² and we take the concatenation of these three sets of results. (a) Univariate logistic regression (LR), performing one-way LR analysis for each characteristic factor and outcome, selecting factors with two-tailed $p \leq 0.05$. (b) Elastic net regression, with regularization methods to optimize model complexity and identify important sparse sets of factors.²³ (c) Random forest, comparing the magnitude of contribution between features, the degree of contribution was measured using the Gini index as an evaluation metric.

The factors screened were then subjected to correlation analysis, using Pearson and Spearman correlation coefficients. Correlation coefficient >0.6 was judged to be strong correlation.

The last step is to conduct a multivariable LR analysis. To avoid multi-collinearity, factors with correlation greater than 0.6 are put into the model separately. The results take the intersection and the factors obtained are included in the next step of the analysis.

Statistical analysis

We performed descriptive analyses of all individuals found in the database who met the study requirements. Mean \pm standard deviation (SD) was used to describe normally distributed continuous factors, and median and IQR (interquartile range) were used to describe skewed continuous factors. Categorical factors were statistically described using frequencies (component ratios). Differences in continuous factors were tested using Student's *t*-test or the rank-sum test. Differences in categorical factors were tested using the Chi-squared test.

In this study, the dataset was randomly divided into a training set and a test set according to the sample size ratio of 7:3. The training set was used to select factors and train the model, and the test set was used to validate the performance of the model. In this study SMOTENC was applied to the training set to address data imbalances hence allowing for better model performance and prediction following past studies.²⁴ By reading the related literatures,^{25,26} considering both the complexity of the model and the interpretability of the results, seven commonly used machine-learning prediction models with high correct rates were selected and covered both machine-learning models and deep learning models. The study used seven methods such as DT, random forests (RF), Gradient Boosted Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), LR, DNN, and Oblique Decision Random Forest (ODRF). The training and testing process involves 5-fold cross-validation. Prediction accuracy, AUC, precision, recall, F1 score, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were then computed and compared to assess the performance of the models and to identify the best model for predicting the risk and prognosis of IS in hypertensive patients. Finally, the best model was interpreted using SHapley Additive exPlanations (SHAP).

The descriptive analysis and factor selection processes were performed using the R programming language (version 4.1.3, R Core Team, Vienna, Austria) and RStudio software (version 2022.02.1, RStudio Team, Boston, MA, USA). The modeling and interpretable learning of the model were performed using the Python programming language (version 3.7, Python Core Development Team, Virginia, USA) and Pycharm software (version 2020.3.4, JetBrains, Czech Republic). The statistical significance level was 0.05 (two-tailed).

Results

Baseline characteristics

The demographic features of 90,768 and 1746 participants in the study of the risk of IS in hypertensive patients who did not suffer from IS and those who suffered from IS, respectively, is shown in Table 1. The mean age was 68.1 ± 14.7 years. The difference in BMI between the two groups was not statistically significant. Hyperlipidemia accounted for the largest proportion of comorbidities at 49.1%, 45.6% for those taking statins, and 42.6% for those taking aspirin.

The number of hypertensive patients suffering from IS prognosis study without death and death were 907 and 839, respectively, and the demographic status are shown in Table 1. The mean age was 75.3 ± 12.5 years. The difference in BMI was not statistically significant between the two groups. The largest proportion of comorbidities was

Table 1. Baseline characteristics of participants.

For Risk Research Populations				
Factors	Total (n = 92,514)	Non-IS (n = 90,768)	IS (n = 1746)	p
Demographic characteristics				
Age, mean \pm SD	68.1 \pm 14.7	68.0 \pm 14.7	72.2 \pm 12.7	<0.001
Gender, female <i>n</i> (%)	44,996 (48.6)	44,061 (48.5)	935 (53.6)	<0.001
Weight, mean \pm SD	183.3 \pm 44.2	183.4 \pm 44.4	180.6 \pm 43.5	0.008
BMI, mean \pm SD	29.1 \pm 5.4	29.1 \pm 5.4	28.9 \pm 5.3	0.351
BMI_category				
1 (<18.5)	1774 (1.9)	1747 (1.9)	27 (1.5)	0.292
2 (18.5–24.9)	19,358 (20.9)	18,981 (20.9)	377 (21.6)	0.507
3 (25.0–29.9)	31,555 (34.1)	30,950 (34.1)	605 (34.7)	0.648
4 (\geq 30.0)	39,827 (43.0)	39,090 (43.1)	737 (42.2)	0.468
Race, <i>n</i> (%)				
Asian, <i>n</i> (%)	2812 (3.0)	2763 (3.0)	49 (2.8)	0.615
Black, <i>n</i> (%)	12,847 (13.9)	12,492 (13.8)	355 (20.3)	<0.001
White, <i>n</i> (%)	62,958 (68.1)	61,812 (68.1)	1146 (65.6)	0.031
Other, <i>n</i> (%)	13,897 (15.0)	13,702 (15.1)	196 (11.2)	<0.001
Comorbidities				
Hyperlipidemia, <i>n</i> (%)	45,384 (49.1)	44,480 (49.0)	904 (0.5)	0.023
Congestive heart failure, <i>n</i> (%)	14,937 (16.1)	14,586 (16.1)	351 (0.2)	<0.001
Peripheral vascular disease, <i>n</i> (%)	7997 (8.6)	7781 (8.6)	216 (12.4)	<0.001
Cerebrovascular disease, <i>n</i> (%)	10,266 (11.1)	10,056 (11.1)	210 (12.0)	0.226
Mild liver disease, <i>n</i> (%)	5780 (6.2)	5705 (6.3)	75 (4.3)	<0.001
Severe liver disease, <i>n</i> (%)	1447 (1.6)	1433 (1.58)	14 (0.8)	0.013
Malignant cancer, <i>n</i> (%)	9598 (10.4)	9422 (10.4)	176 (10.1)	0.713
Myocardial infarct, <i>n</i> (%)	10,490 (11.3)	10,261 (11.3)	229 (13.1)	0.020
Renal disease, <i>n</i> (%)	15,710 (17.0)	15,293 (16.8)	417 (23.8)	<0.001
Chronic pulmonary disease, <i>n</i> (%)	17,916 (19.4)	17,604 (19.4)	312 (17.9)	0.117
Sepsis, <i>n</i> (%)	3708 (4.0)	3679 (4.1)	29 (1.7)	<0.001
Cerebrovascular atherosclerosis, <i>n</i> (%)	548 (0.6)	535 (0.6)	13 (0.7)	0.497

For Prognosis Research Populations				
	Total (<i>n</i> = 1746)	Nondeath (<i>n</i> = 907)	Death (<i>n</i> = 839)	<i>p</i>
Demographic characteristics				
Age, mean ± SD	75.3 ± 12.5	72.6 ± 12.3	78.3 ± 12.0	<0.001
Gender, female <i>n</i> (%)	935 (53.6)	467 (51.5)	468 (55.8)	0.072
Weight, mean ± SD	178.3 ± 42.0	182.4 ± 41.1	173.8 ± 42.5	<0.001
BMI, mean ± SD	28.8 ± 5.0	29.0 ± 4.9	28.7 ± 5.2	0.313
BMI_category				
1 (<18.5)	23 (1.3)	12 (1.3)	11 (1.3)	1.0
2 (18.5–24.9)	375 (21.5)	174 (19.2)	201 (24.0)	0.018
3 (25.0–29.9)	635 (36.4)	339 (37.4)	296 (35.3)	0.390
4 (≥30.0)	713 (40.8)	383 (42.2)	330 (39.3)	0.238
Race, <i>n</i> (%)				
Asian, <i>n</i> (%)	52 (3.0)	24 (2.6)	28 (3.3)	0.474
Black, <i>n</i> (%)	351 (20.1)	188 (20.7)	163 (19.4)	0.553
White, <i>n</i> (%)	1184 (67.8)	600 (66.2)	582 (69.4)	0.146
Other, <i>n</i> (%)	159 (9.1)	96 (10.6)	65 (7.7)	0.051
Comorbidities				
Hyperlipidemia, <i>n</i> (%)	1093 (62.6)	594 (65.5)	499 (59.5)	0.013
Congestive heart failure, <i>n</i> (%)	569 (32.6)	240 (26.5)	329 (39.2)	<0.001
Peripheral vascular disease, <i>n</i> (%)	275 (15.8)	138 (15.2)	137 (16.3)	0.553
Mild liver disease, <i>n</i> (%)	87 (5.0)	38 (4.2)	49 (5.8)	0.138
Severe liver disease, <i>n</i> (%)	27 (1.5)	3 (0.3)	24 (2.9)	<0.001
Malignant cancer, <i>n</i> (%)	222 (12.7)	65 (7.2)	157 (18.7)	<0.001
Myocardial infarct, <i>n</i> (%)	323 (18.5)	150 (16.5)	173 (20.6)	0.031
Renal disease, <i>n</i> (%)	571 (32.7)	232 (25.6)	339 (40.4)	<0.001
Chronic pulmonary disease, <i>n</i> (%)	354 (20.3)	178 (19.6)	176 (21.0)	0.505
Sepsis, <i>n</i> (%)	123 (7.0)	29 (3.2)	94 (11.2)	<0.001
Cerebrovascular atherosclerosis, <i>n</i> (%)	117 (6.7)	74 (8.2)	43 (5.1)	0.015

(continued)

Table 1. Continued.

For Prognosis Research Populations				
	Total (<i>n</i> = 1746)	Nondeath (<i>n</i> = 907)	Death (<i>n</i> = 839)	<i>p</i>
Coagulopathy, <i>n</i> (%)	150 (8.6)	56 (6.2)	94 (11.2)	<0.001
Respiratory failure, <i>n</i> (%)	160 (9.2)	45 (5.0)	115 (13.7)	<0.001

hyperlipidemia at 62.6%, 74.6% of those taking aspirin, and 67.4% of those taking statins. Descriptive analyses of factors in the laboratory results and medicine section in Supplementary Table S1.

Factor selection

For the study of the risk of IS, the predictor factors that were closely related were age, sex, peripheral vascular disease, renal disease, aspirin, glucose maximum, COPD, hyperlipidemia, amlodipine, sepsis, neutrophil maximum, statins, and BMI. Predictor factors that were strongly associated with prognostic studies of IS were age, CCI, RDW maximum, BMI, sepsis, triglyceride, rivaroxaban, respiratory failure, statins, and acetaminophen.

Model development and validation

Because of the imbalance in the proportion of hypertensive patients with and without, the SMOTENC balancing technique was applied to the training dataset before modeling. We applied machine-learning algorithms with the training dataset and validated the model using the test

dataset (parameters settings are shown in Supplementary Table S2).

The prediction results of the seven machine-learning algorithms are shown in Supplementary Table S3. The comparison of AUC of the seven machine-learning algorithms is shown in Figure 1A and B. From Supplementary Table S3, it can be seen that the model with the highest accuracy and AUC is the GBDT model.

Model interpretation

Our study found that the GBDT model performed the best in risk prediction and prognosis prediction.

The basic idea of the GBDT is to combine many weak base classifiers into one strong base classifier.²⁷ The advantages of GBDT are good training results, less overfitting, and flexibility in handling various data types, including continuous and discrete values.²⁸ GBDT is a model with strong generalization capabilities. A number of medical studies have used the GBDT model.²⁹

We then performed a SHAP analysis of the GBDT model to reveal the distribution of the effects of each selected factor.

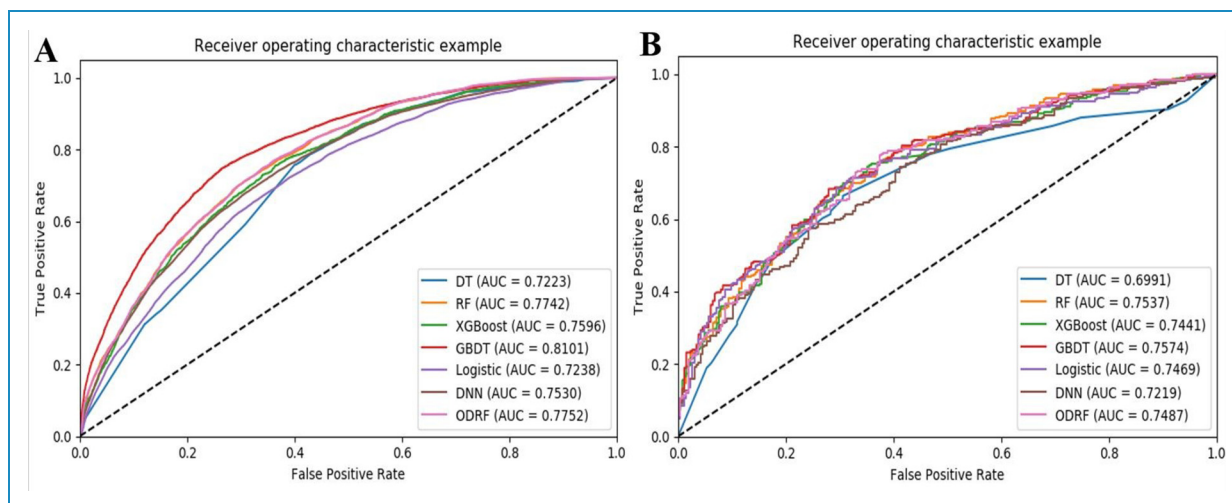


Figure 1. Performance evaluation of seven machine-learning algorithms with ROC curves. (A) ROC curves of seven models for predicting risk of stroke in hypertensive patients. (B) ROC curves of seven models for predicting prognosis of stroke in hypertensive patients.

Firstly, in order to determine the importance of each feature to the predictive model, the SHAP summary of the GBDT model was plotted Figure 2.

Red dots indicate high-risk value and blue indicate low.³⁰ As shown in Figure 2A and B, high values of age, BMI, neutrophil maximum, and glucose maximum correspond to SHAP values greater than zero. This suggests that these characteristics are important risk factors.

As shown in Figure 2C and D, high values for age, CCI, RDW maximum, and triglyceride corresponded to SHAP values greater than zero. This suggests that these features are important factors for the prognosis.

Secondly, the practical application of the model takes the form shown in Supplementary Figure S2. Red areas indicate that the eigenvalue increases the probability of the ending occurring, while blue areas indicate that the eigenvalue decreases. $f(x)$ represents the composite SHAP value for each patient. The mean SHAP value of all samples was used as the baseline value. When the value of $f(x)$ is greater than the baseline value, the model predicts that the ending occurs.³¹ Supplementary Figure S2A shows that the hypertensive patient got an IS and the model predicted true. Supplementary Figure S2B shows that the hypertensive patient did not get an IS and the model predicted true. Supplementary Figure S2C shows that the patient died and the model predicted true. Supplementary Figure S2D shows that the patient did not die and the model accurately predicted true.

The importance ranking of factor interactions was plotted as shown in Figure 3A and B. The interactions between age and BMI, glucose maximum and aspirin, age

and glucose maximum, and sex and age were significant for IS risk prediction. As shown in Figure 3B, the interactions between age and CCI, age and RDW maximum, age and respiratory failure, CCI and triglyceride, age and triglyceride, CCI and RDW maximum were significant for the prognostic prediction.

As shown in Figure 4A, with increasing age, the risk of IS is higher in hypertensive patients with higher BMI, especially those between 40 and 90 years of age.

As shown in Figure 4B, for hypertensive patients with higher than normal blood glucose (fasting blood glucose normal values of 3.9–6.1 mmol/L, database in mg/dL, 6.1 mmol/L is approximately equal to 109.8 mg/dL). There are more red dots below the SHAP value = 0 than above the SHAP value = 0, indicating that taking aspirin is more effective in reducing the risk of IS in hypertensive patients with high blood glucose values.

As shown in Figure 4C, with increasing age, blood glucose values outside the normal range increase the risk, especially those between the ages of 40 and 90 years old.

As shown in Figure 4D females (gender = 0) should take care of IS prevention after 60 years of age, and males (gender = 1) should prevent IS when they are around 45 years of age.

As shown in Figure 4E the CCI increases with age and the higher age CCI is detrimental to the prognosis.

The normal range of RDW is 11.5% to 14.5%, as shown in Figure 4F high RDW values are detrimental to the prognosis as age increases.

As shown in Figure 4G with increasing age, having respiratory failure disease is detrimental to the prognosis.

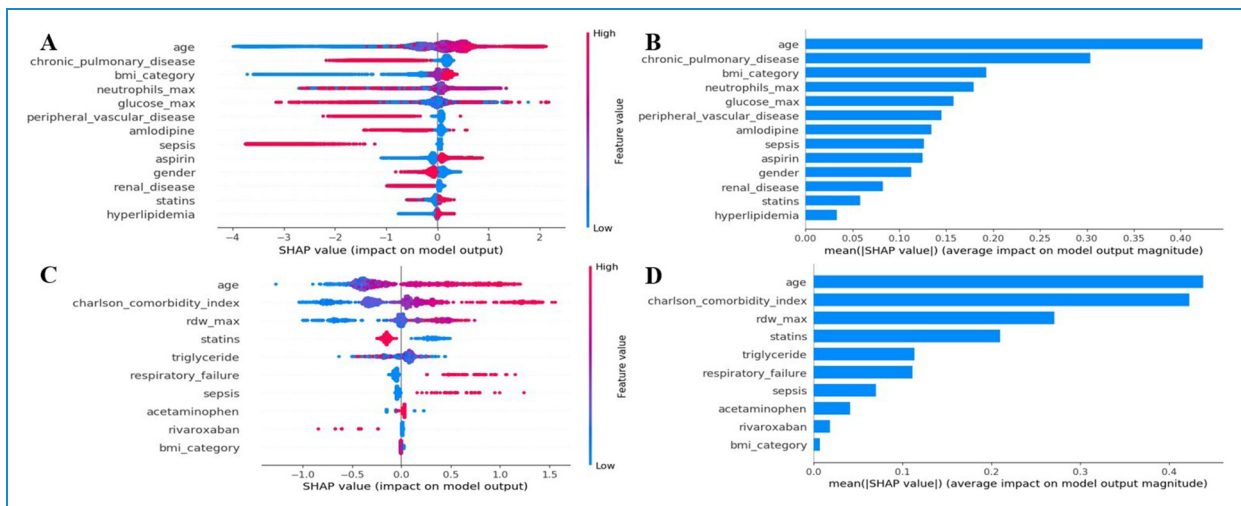


Figure 2. Summary of SHAP for the GBDT model. (A) The higher the characteristic SHAP value the more likely IS is to occur. A point is created in the model to represent one feature attribute value for one patient, so a point is assigned to each feature on the line for each patient. Points are colored according to the feature values of the corresponding patient and accumulated vertically to depict density. Red color indicates higher feature values and blue color indicates lower feature values. (B) The absolute value of the mean of the SHAP values for each feature is the feature importance distribution. (C) The higher the SHAP value of a feature, the more likely it is that a poor prognosis for IS will occur. (D) The absolute value of the mean of the SHAP values for each feature is the feature importance distribution.

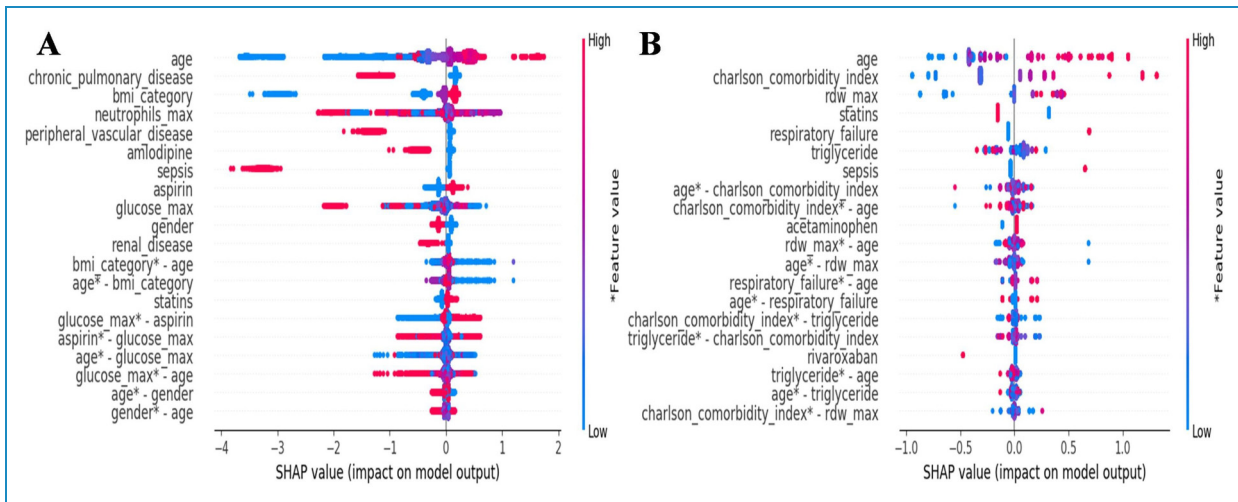


Figure 3. Ranking of the importance of variable interactions. (A) Important risk factors in the predictive model of the risk of ischemic stroke in hypertensive patients. (B) Important risk factors in the predictive model of the prognosis of ischemic stroke in hypertensive patients.

The normal range for triglyceride is 0.45 to 1.69 mmol/L, database unit is mg/dL, unit conversion $\text{mg/dL} * 0.011 = \text{mmol/L}$, and exceeding 149.7 mg/dL is outside the normal range. As shown in Figure 4H, more and more patients had triglycerides outside the normal values with increasing CCI. Both patient comorbidity with other diseases and triglycerides exceeding normal values are detrimental to the prognosis.

As shown in Figure 4I as age increases, the more triglycerides exceed normal values the less favorable the prognosis.

As shown in Figure 4J as the CCI increases, the number of patients with RDW exceeding normal values is increasing, and having other comorbidities with RDW exceeding normal values is detrimental to the prognosis.

Discussion

Over the past few decades, the global burden of stroke has increased dramatically, especially in low- and middle-income countries, because of the increasing in aging populations and modifiable stroke risk factors.³ The correlation between hypertension and increased risk of stroke has long been the strongest and most recognized.³²

Our study, in line with other related studies, identified a number of influential factors that affect the risk and prognosis of IS in hypertensive patients. The importance of our study lies in the fact that we used statistical methods and machine-learning models to determine the impact on the onset and prognosis IS in patients with hypertension and identified specific high-risk subgroups with above-average responses to specific risk factors.

Data imbalance is a common problem in many real-world datasets, which can seriously affect machine-learning model performance, as many models are sensitive to the distribution of classes, so we perform data balancing on

the training set. Feature importance is an important aspect of understanding the predictive power of a model. The GBDT model provides built-in methods for calculating feature importance, which can help us understand which features have the most impact on prediction. Combining all the result metrics reveals that the GBDT model achieves the best performance on the internal validation set compared to the other models. We used the SHAP method to perform interpretable learning on the GBDT model to explore the risk and prognostic influences affecting the risk of IS in hypertensive patients.

Both global and regional changes in brain tissue volume occur specifically with age.³³ Aging is the most important factor influencing the incidence and prevalence of stroke.³⁴ The incidence of IS increases with age, especially for those aged 50 to 69 years or older.³⁵ Although age is un-modifiable, it is important to pay more attention to stroke prevention as age increases.

Previously published study has suggested that risk of IS is higher for those with obesity,³⁶ high blood glucose,³⁷ while RDW is associated with poor prognosis³⁸ and higher triglyceride indices are associated with a higher risk of poor functional prognosis and in-hospital mortality.³⁹ Our findings also verified these findings, which also suggested the reliability of our analysis.

Published study found that neutrophils are the first cells in the peripheral blood to reach the infarcted area of the brain after the onset of IS.⁴⁰ Previous studies have found that the number of neutrophils in the area of cerebral infarction increases over time.⁴¹ High levels of neutrophil count and neutrophil ratio were found to be associated with mild IS or transient ischemic attack, as well as an increased risk of IS.⁴² Similarly, our study found that high levels of neutrophil ratio can be a risk factor and can be used as a diagnostic aspect for clinicians.

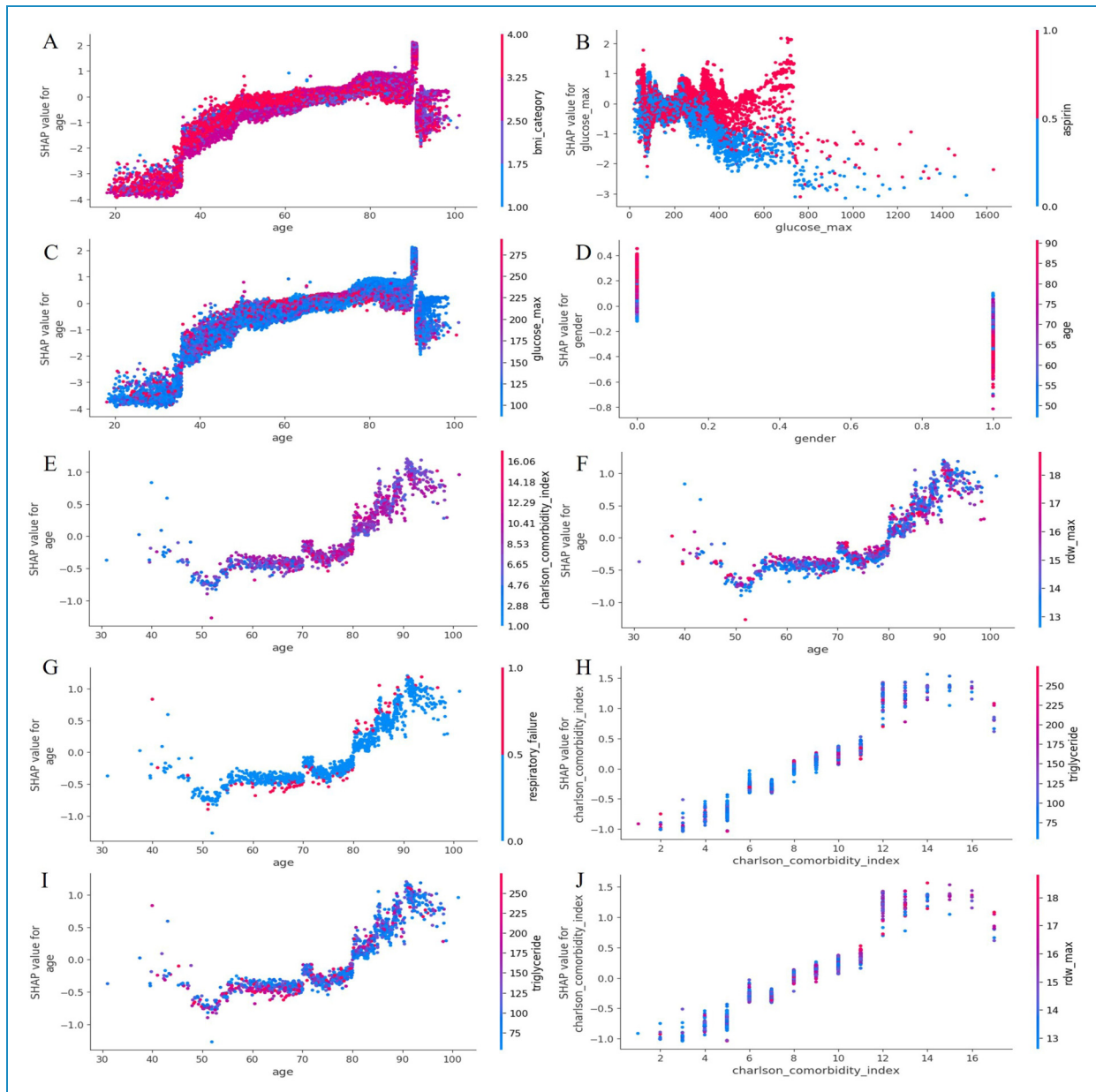


Figure 4. Variable interaction dependency plots. (A) Interaction between age and BMI on the risk of having an ischemic stroke. (B) Interaction between blood glucose maximum and aspirin on the risk of having an ischemic stroke. (C) Interaction between age and blood glucose maximum on the risk of having an ischemic stroke. (D) Interaction between sex and age on the risk of having an ischemic stroke. (E) Interaction between age and the CCI on the prognosis of having an ischemic stroke. (F) Interaction between age and RDW maximum on the prognosis of having an ischemic stroke. (G) Interaction between age and respiratory failure on the prognosis of having an ischemic stroke. (H) Interaction between the CCI and triglyceride on the prognosis of having an ischemic stroke. (I) Interaction between age and triglyceride on the prognosis of having an ischemic stroke. (J) Interaction between CCI and RDW maximum on prognosis of having ischemic stroke.

According to our results, comorbidities are an unfavorable prognostic factor. Treatment after the first IS occurs should be accompanied by attention to the treatment and prevention of other comorbidities.

The relationship between factors affecting hypertensive patients suffering from IS are complex, the effect of each

factor is also simultaneously influenced by the others. Statistically, interaction is representative of the moderating effect between factors.⁴³ In order to better understand the relationship of factors associated with IS risk and prognosis for hypertensive patients, we also conducted an interaction effect analysis.

Gender differences of IS depend on the age of the patient with hypertension, as the impact of gender on stroke risk and prognosis changes throughout the life cycle.⁴⁴ This is consistent with our findings in the interaction of sex and age on the risk of developing IS. In middle age, our findings suggested that more attention should be given to the prevention of IS in men. While in older age, the incidence of IS begins to increase in women. This may be related to hormones. A large meta-analysis reports that the risk of stroke associated with metabolic syndrome is significantly higher in women than in men.⁴⁵ One study found that women with early menopause had a twofold increased risk of IS.⁴⁶

The interaction of age and BMI also influenced the risk. Changes in important risk factors for thromboembolic stroke during aging may be related to weight gain, with particular attention to elevate blood pressure or the acute phase of hypertension.⁴⁷ Therefore, it is important to focus on weight management for older hypertensive patients.

In the analysis, interaction between aspirin and blood glucose was significant. Hyperglycemia increases the size of cerebral infarcts and the permeability of the blood-brain barrier,⁴⁸ so hypertensive patients need to pay attention to glycemic control in order to reduce the risk.⁴⁹ Meanwhile, antiplatelet drugs can be used to reduce the long-term risk of non-cardiogenic embolic IS,⁵⁰ so hypertensive patients can take aspirin reasonably to prevent IS.

The interaction effect of age and blood glucose was also significant, suggesting that the older the more important it is to control blood glucose in hypertensive patients.

Increased mortality in the elderly may be related to the effects of aging and late complications.³³ Results of a mixed-gender clinical study showed older stroke patients appeared to be more likely to develop infectious complications, such as pneumonia and urinary tract infections.⁵¹ In our study the interactions of age and CCI, as well as the interaction between age and respiratory failure were found to have some effect on the prognosis.

The interactions of CCI and RDW, as well as age and RDW also affected the prognosis as suggested by our findings. A growing number of studies have shown that RDW is strongly associated with the incidence and prognosis of many diseases, such as myocardial infarction,⁵² and so on. The higher the CCI is, the less favorable the patient's prognosis.⁵³

Studies have shown that both inflammatory response and oxidative stress are associated with RDW.⁵⁴ Oxidative stress affects the life span of red blood cell, destroys the red blood cell membrane, and increases the osmotic fragility of red blood cell as well as their ability to adhere and aggregate.⁵⁴ All these factors lead to changes in RDW. In addition, these pathological processes can promote coagulation and thrombosis.⁵⁵ Meanwhile, several studies have identified potential mechanisms for the effects of aging, which include the promotion of oxidative stress and inflammatory responses.⁵⁶ Both RDW and age are relatively

simple and readily available indicators that can help physicians better identify and assess the prognosis.

In our study, we found that the interaction between CCI and triglyceride, as well as the interaction of age and triglyceride also influence the prognosis. Several studies have shown that plasma triglyceride concentrations accumulate with age.⁵⁷ Triglyceride accumulation leads to the formation of atherosclerotic plaques and are also a risk factor for diseases including peripheral arterial disease,⁵⁸ and so on. Control of triglyceride reduces the incidence of comorbid diseases and benefits the prognosis.

There are also some limitations to this study. Firstly, our study was a retrospective research conducted based on public database, thus further prospective study is needed to verify our findings. Then, several important measurements in IS research such as the NIHSS scores, results of cranial CT and MRI are not available in the MIMIC-IV database, which may cause the loss of information and potential bias in the analysis. Besides, data missing occurs for almost all factors, though we have conducted missing data imputation to minimize the influence, potential bias may not be totally avoided.

Conclusions

Combining all the resultant metrics, GBDT model obtained the best performance on the internal validation set as compared to the other models. Subgroup identification analysis suggests that hypertensive patients with specific characteristics have a higher risk of IS and a poor prognosis. Our findings may provide a reference for the prevention of IS and the improvement of prognosis after disease onset in the hypertensive population.

Availability of data and materials: The datasets generated and/or analyzed during the current study can be found in MIMIC-IV (<https://physionet.org/content/mimiciv/2.0/>).

Contributors: F.C. and J.M. contributed to the conception of the study, supervised the analysis, and conducted critical revision of the manuscript. L.G. conducted data curation and management, the formal analysis, the presentation of results and graphs, drafted the original manuscript, and the revision of the manuscript. S.C. and Y.Y. conducted data curation and the draft of the manuscript. W.H., J.C., and S.L. assisted in the analysis. Y.Z. and L.P. assisted in the revision of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Not applicable, because this article does not contain any studies with human or animal subjects.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this

article: This work was supported by the National Social Science Fund of China (21CTJ009), the Natural Science Basic Research Program of Shaanxi Province, China (2022JQ-769), and the National Natural Science Foundation of China (81703325).

Guarantor: FC.

ORCID iD: Fangyao Chen  <https://orcid.org/0000-0002-7433-9424>

Supplemental material: Supplemental material for this article is available online.

References

1. Feigin VL and Owolabi MO, World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group. Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission. *Lancet Neurol* 2023; 22: 1160–1206. Epub 2023 Oct 9. Erratum in: *Lancet Neurol*. 2023 Dec; 22(12): e13. doi: 10.1016/S1474-4422(23)00425-8.
2. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol* 2024; 23: 344–381. Epub 2024 Mar 14. Erratum in: *Lancet Neurol*. 2024 May; 23(5): e9. doi: 10.1016/S1474-4422(24)00114-5. Erratum in: *Lancet Neurol*. 2024 Jul; 23(7): e11. doi: 10.1016/S1474-4422(24)00231-X.
3. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. *Int J Stroke* 2022; 17: 18–29.
4. Pu L, Wang L, Zhang R, et al. Projected global trends in ischemic stroke incidence, deaths and disability-adjusted life years from 2020 to 2030. *Stroke* 2023; 54: 1330–1339. Epub 2023 Apr 24. Erratum in: *Stroke*. 2024 Jan; 55(1): e23.
5. Malikova H and Weichet J. Diagnosis of ischemic stroke: as simple as possible. *Diagnostics (Basel)* 2022; 12: 1452.
6. Ding Y, Lang Y, Zhang H, et al. Candesartan reduces neuronal apoptosis caused by ischemic stroke via regulating the FFAR1/ITGA4 pathway. *Mediators Inflamm* 2022; 2022: 2356507.
7. Liu S, Li Y, Zeng X, et al. Burden of cardiovascular diseases in China, 1990–2016: findings from the 2016 global burden of disease study. *JAMA Cardiol* 2019; 4: 342–352.
8. Wang YJ, Li ZX, Gu HQ, et al. China stroke statistics: an update on the 2019 report from the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. *Stroke Vasc Neurol* 2022; 7: 415–450.
9. Cho H, Kim T, Koo J, et al. Untreated hypertension and prognosis paradox in acute ischemic stroke. *Neurol Sci* 2023; 44: 2087–2095.
10. Abujaber AA, Alkhalil IM, Imam Y, et al. Predicting 90-day prognosis for patients with stroke: a machine learning approach. *Front Neurol* 2023; 14: 1270767.
11. Abujaber AA, Albalkhi I, Imam Y, et al. Machine learning-based prognostication of mortality in stroke patients. *Heliyon* 2024; 10: e28869.
12. Huang X and Zhang Y. Relationship between serum bicarbonate levels and the risk of death within 30 days in ICU patients with acute ischemic stroke. *Front Neurol* 2023; 14: 1125359.
13. Cai W, Xu J, Wu X, et al. Association between triglyceride-glucose index and all-cause mortality in critically ill patients with ischemic stroke: analysis of the MIMIC-IV database. *Cardiovasc Diabetol* 2023; 22: 138.
14. Li H, Liu P, Ma HY, et al. Novel predictors and a predictive model of cerebrovascular atherosclerotic ischemic stroke based on clinical databases. *Neurol Res* 2023; 45: 391–399.
15. Jin G, Hu W, Zeng L, et al. Prediction of long-term mortality in patients with ischemic stroke based on clinical characteristics on the first day of ICU admission: an easy-to-use nomogram. *Front Neurol* 2023; 14: 1148185.
16. Zheng Y, Guo Z, Zhang Y, et al. Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine. *EPMA J* 2022; 13: 285–298.
17. Liu F and McCullough LD. Interactions between age, sex, and hormones in experimental ischemic stroke. *Neurochem Int* 2012; 61: 1255–1265.
18. Horn JW, Feng T, Mørkedal B, et al. Body mass index measured repeatedly over 42 years as a risk factor for ischemic stroke: the HUNT study. *Nutrients* 2023; 15: 1232.
19. Liu Q, Wu S, Shao J, et al. Metabolic syndrome parameters' variability and stroke incidence in hypertensive patients: evidence from a functional community cohort. *Cardiovasc Diabetol* 2024; 23: 3.
20. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023; 10: 1.
21. Lange S, Helfrich K and Ye Q. Batch normalization preconditioning for neural network training. *J Mach Learn Res* 2022; 23: 1. Article 72 (January 2022), 41 pages.
22. Chen S, Hu W, Yang Y, et al. Predicting six-month re-admission risk in heart failure patients using multiple machine learning methods: a study based on the Chinese heart failure population database. *J Clin Med* 2023; 12: 70.
23. Munch MM, Peeters CFW, Van Der Vaart AADW, et al. Adaptive group-regularized logistic elastic net regression. *Biostatistics* 2021; 22: 723–737.
24. Zhang X, Fei N, Zhang X, et al. Machine learning prediction models for postoperative stroke in elderly patients: analyses of the MIMIC database. *Front Aging Neurosci* 2022; 14: 897611.
25. Chang HW, Zhang H, Shi GP, et al. Ischemic stroke prediction using machine learning in elderly Chinese population: the Rugao Longitudinal Ageing Study. *Brain Behav* 2023; 13: e3307.
26. Hu W, Jin T, Pan Z, et al. An interpretable ensemble learning model facilitates early risk stratification of ischemic stroke in intensive care unit: development and external validation of ICU-ISPM. *Comput Biol Med* 2023; 166: 107577.
27. Rao H, Shi X, Rodrigue AK, et al. Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl Soft Comput* 2018; 74: 634–642.

28. Zhang Y, Zhang R, Ma Q, et al. A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Trans* 2020; 100: 210–220.
29. Rauf A, Ullah A, Rathi U, et al. Predicting stroke and mortality in mitral stenosis with atrial flutter: a machine learning approach. *Ann Noninvasive Electrocardiol* 2023; 28: e13078.
30. Wang K, Tian J, Zheng C, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput Biol Med* 2021; 137: 104813.
31. Li W, Song Y, Chen K, et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *BMJ Open* 2021; 11: e050989.
32. Ferrari F and Villa RF. Brain bioenergetics in chronic hypertension: risk factor for acute ischemic stroke. *Biochem Pharmacol* 2022; 205: 115260.
33. Bretzner M, Bonkhoff AK, Schirmer MD, et al. Radiomics-derived brain age predicts functional outcome after acute ischemic stroke. *Neurology* 2023; 100: e822–e833.
34. Simmons CA, Poupore N and Nathaniel TI. Age stratification and stroke severity in the telestroke network. *J Clin Med* 2023; 12: 1519.
35. Ding Q, Liu S, Yao Y, et al. Global, regional, and national burden of ischemic stroke, 1990–2019. *Neurology* 2022; 98: e279–e290.
36. Harshfield EL, Georgakis MK, Malik R, et al. Modifiable lifestyle factors and risk of stroke: a Mendelian randomization analysis. *Stroke* 2021; 52: 931–936.
37. Huang YQ, Lo K, Liu XC, et al. The relationship between fasting blood glucose levels and first ischemic stroke in elderly hypertensive patients. *Risk Manag Healthc Policy* 2020; 13: 777–784.
38. Fan H, Liu X, Li S, et al. High red blood cell distribution width levels could increase the risk of hemorrhagic transformation after intravenous thrombolysis in acute ischemic stroke patients. *Aging (Albany NY)* 2021; 13: 20762–20773.
39. Miao M, Bi Y, Hao L, et al. Triglyceride-glucose index and short-term functional outcome and in-hospital mortality in patients with ischemic stroke. *Nutr Metab Cardiovasc Dis* 2023; 33: 399–407.
40. Sun S, Lv W, Li S, et al. Smart liposomal nanocarrier enhanced the treatment of ischemic stroke through neutrophil extracellular traps and cyclic guanosine monophosphate-adenosine monophosphate synthase-stimulator of interferon genes (cGAS-STING) pathway inhibition of ischemic penumbra. *ACS Nano* 2023; 17: 17845–17857.
41. Denorme F, Portier I, Rustad JL, et al. Neutrophil extracellular traps regulate ischemic stroke brain injury. *J Clin Invest* 2022; 132: e154225.
42. Bui TA, Jickling GC and Winship IR. Neutrophil dynamics and inflammaging in acute ischemic stroke: a transcriptomic review. *Front Aging Neurosci* 2022; 14: 1041333.
43. Bailly A, Blanc C, Francis É, et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Comput Methods Programs Biomed* 2022; 213: 106504.
44. Roy-O'Reilly M and McCullough LD. Age and sex are critical factors in ischemic stroke pathology. *Endocrinology* 2018; 159: 3120–3131.
45. Alipour P, Azizi Z, Raparelli V, et al. Role of sex and gender-related variables in development of metabolic syndrome: a prospective cohort study. *Eur J Intern Med* 2024; 121: 63–75.
46. Ingelsson E, Lundholm C, Johansson AL, et al. Hysterectomy and risk of cardiovascular disease: a population-based cohort study. *Eur Heart J* 2011; 32: 745–750.
47. Fan J, Li X, Yu X, et al. Global burden, risk factor analysis, and prediction study of ischemic stroke, 1990–2030. *Neurology* 2023; 101: e137–e150.
48. Sifat AE, Nozohouri S, Archie SR, et al. Brain energy metabolism in ischemic stroke: effects of smoking and diabetes. *Int J Mol Sci* 2022; 23: 8512.
49. Denorme F, Portier I, Kosaka Y, et al. Hyperglycemia exacerbates ischemic stroke outcome independent of platelet glucose uptake. *J Thromb Haemost* 2021; 19: 536–546.
50. Lin MP, Meschia JF, Gopal N, et al. Cilostazol versus aspirin for secondary stroke prevention: systematic review and meta-analysis. *J Stroke Cerebrovasc Dis* 2021; 30: 105581.
51. Ahmed R, Mhina C, Philip K, et al. Age- and sex-specific trends in medical complications after acute ischemic stroke in the United States. *Neurology* 2023; 100: e1282–e1295.
52. de Albuquerque FVDS C, Dias-Neto MF, Rocha-Neves JMPD, et al. Red blood cell distribution width predicts myocardial infarction and mortality after vascular surgery—a prospective cohort study. *World J Surg* 2022; 46: 957–965.
53. Barow E, Probst AC, Pinnschmidt H, et al. Effect of comorbidity burden and polypharmacy on poor functional outcome in acute ischemic stroke. *Clin Neuroradiol* 2023; 33: 147–154.
54. Vo HH, Truong-Thi NN, Ho-Thi HB, et al. The value of neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, red cell distribution width, and their combination in predicting acute pancreatitis severity. *Eur Rev Med Pharmacol Sci* 2023; 27: 11464–11471.
55. Ling J, Fang M and Wu Y. Association of red cell distribution width and D-dimer levels with intracranial hemorrhage in patients with cerebral venous thrombosis. *Clin Neurol Neurosurg* 2022; 214: 107178.
56. Mozzini C and Pagani M. Oxidative stress in chronic and age-related diseases. *Antioxidants (Basel)* 2022; 11: 21.
57. Spitzer KM and Davies BSJ. Aging and plasma triglyceride metabolism. *J Lipid Res* 2020; 61: 1161–1167.
58. Gao JW, Hao QY, Gao M, et al. Triglyceride-glucose index in the development of peripheral artery disease: findings from the Atherosclerosis Risk in Communities (ARIC) study. *Cardiovasc Diabetol* 2021; 20: 26.