Journal of Clinical and Nursing Research, 2023, Volume 7, Issue 3
http://ojs.bbwpublisher.com/index.php/JCNR
ISSN Online: 2208-3693
ISSN Print: 2208-3685

An Interpretable Prediction Model for Stroke Based on XGBoost and SHAP

Tianshu Fang[1,2], Jiacheng Deng[2]*

[1]Nanchang University Queen Mary School, Nanchang University, Nanchang 330031, Jiangxi Province, China

[2]Nanchang Key Laboratory of Medical and Technology Research, Nanchang University, Nanchang 330031, Jiangxi Province, China

***Corresponding author:** Jiacheng Deng, tgmpg183@163.com

Abstract: *Objective:* To establish a stroke prediction and feature analysis model integrating XGBoost and SHAP to aid the clinical diagnosis and prevention of stroke. *Methods:* Based on the open data set on Kaggle, with the help of data preprocessing and grid parameter optimization, an interpretable stroke risk prediction model was established by integrating XGBoost and SHAP and an explanatory analysis of risk factors was performed. *Results:* The XGBoost model's accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC) were 96.71%, 93.83%, 99.59%, and 99.19%, respectively. Our explanatory analysis showed that age, type of residence, and history of hypertension were key factors affecting the incidence of stroke. *Conclusion:* Based on the data set, our analysis showed that the established model can be used to identify stroke, and our explanatory analysis based on SHAP increases the transparency of the model and facilitates medical practitioners to analyze the reliability of the model.

Keywords: Stroke risk prediction; XGBoost algorithm; SHAP model; Risk factor analysis

Online publication: May 30, 2023

1. Introduction

Stroke can be classified as ischemic stroke or hemorrhagic stroke. Due to its high morbidity and mortality, the health and safety of our citizens are at risk. According to the Global Burden of Disease (GBD), the cerebrovascular mortality rate of Chinese residents in 2018 was 149/100,000 (the death toll was 1.57 million), an increase in 41% compared with that in 1990; it is the third cause of death among urban residents and the second cause of death among rural residents [1]. However, among the more than 90% of the stroke burden caused by potentially modifiable risk factors, more than 75% of the stroke burden can be reduced by controlling metabolic and behavioral risk factors [2]. Therefore, it is of great significance to establish an accurate stroke risk prediction model and mine the key risk factors of stroke from the data.

The early stroke risk assessment focused on influencing factors, and statistical methods such as systematic review, meta-analysis, Cox proportional hazards regression model, and logistic regression model were used to identify key risk factors [3]. In recent years, machine learning algorithms have played an important role in the field of disease prediction. Through the establishment of a machine learning risk prediction model to identify stroke accurately and scientifically as well as provide an explanation for the model based on the inherent laws of the data, medical practitioners will be able to make better decisions.

2. Literature review

As the incidence of stroke continues to rise, along with the economic burden of treatment, the scientific identification of risk factors and the early and accurate prediction of individual incidence have become the top priorities of stroke prevention and treatment. Researchers in the field have applied statistics, machine learning, and deep learning methods to their research work (**Table 1**) in order to identify and analyze stroke-related risk factors as well as provide support for the primordial prevention and primary prevention of stroke. Barra *et al*. [4] modified the R-CHA2DS2VASc score, verified its predictive performance on 1,711 patients with myocardial infarction, and obtained the improved R-CHA2DS2VASc score to predict the area under the ROC curve (AUC; 0.717) of ischemic stroke. Based on three different 10-year cohort follow-ups of random population samples from three regions in Finland, Vartiainen *et al*. [5] proposed the FINRISK calculator, which is now widely used to predict the risk of coronary heart disease, stroke incidence, and its combination. Hou *et al*. [6] applied logistic regression to extract feature variables, divided 1,288 effective samples into two sets (a training set and a test set) in an 8:2 ratio, and trained with support vector machine algorithm to obtain a complex clinical-level model with an accuracy rate of 91.09% on the test set. Luo *et al*. [7] used the current parameters of their patients and those of the regression prediction one year later as input data, as well as adopted the bidirectional long-short term memory (BiLSTM) model with Attention mechanism. After training on a data set containing 983 samples, the accuracy rate of the test set was 86%. Chen *et al*. [8] proposed a novel stroke risk prediction framework based on deep transfer learning for data imbalance and small amount of data dispersion, Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP), which has been extensively tested in synthetic and real-world scenarios.

Table 1. Comparison of existing studies

Researchers	Year	Risk factors	Research methods
Barra *et al*. [4]	2013	Age, gender, acute myocardial infarction, diabetes, history of stroke, hypertension, coronary artery disease, peripheral arterial disease, glomerular filtration rate, blood urea nitrogen, and mean $CHADS_2$ score	Modified R-CHA_2DS_sVASc score
Vartiainen *et al*. [5]	2016	History of smoking, systolic blood pressure, total cholesterol, high-density lipoprotein, diabetes, and family history of stroke	FINRISK calculator
Hou *et al*. [6]	2019	Gender, occupation, and history of heart disease, hypertension, diabetes, smoking, hyperhomocysteinemia, hyperlipidemia, hyperuricemia, atherosclerosis, and carotid artery stenosis	Support vector machines
Luo *et al*. [7]	2021	Gender, age, history of hypertension, C-reactive protein, total cholesterol, LPA, arterial plaque, degree of stenosis, and intima-media thickness	BiLSTM model with Attention mechanism
Chen et al. [8]	2022	Age, neutrophil count, lymphocyte count, eosinophil count, basophil count, total protein, albumin, globulin, total bilirubin, direct bilirubin, blood urea nitrogen, *etc*.	Deep transfer learning

Most of the studies above are based on small data sample training models [4,6,7], and the accuracy of the test sets is also average [6,7,]. Foreign research mainly focused on American and European populations [5], which may not be applicable to our country. Although the model based on deep transfer learning [8] can solve the overfitting problem of small data sets, the interpretability of the model is poor. In this study, we used the public data set on Kaggle, which contains a total of 5,110 valid clinical samples, and XGBoost (eXtreme Gradient Boosting) early stroke risk prediction model to assess individual stroke risk. The algorithm has good application advantages, as it excludes complex test indicators and uses only common

demographic data, clinical history, and lifestyle factors related to stroke. At the same time, it integrates SHAP (SHapley Additive exPlanations), a unified framework for explaining predictions. SHAP analyzes the contribution of different risk factors to stroke risk, so that clinicians can further conduct comprehensive analyses of stroke risk prediction models.

3. Methods

The process of constructing the interpretable stroke prediction model combined with XGBoost and SHAP is shown in **Figure 1**. First, data preprocessing such as outlier processing, feature variable coding, and data imbalance processing was performed on the original data; second, the XGBoost risk prediction model was constructed, and its hyperparameters were optimized; third, performance comparison analysis was carried out with the models constructed by different algorithms to verify the prediction effect of the model; lastly, an explanatory analysis of the risk factors in the prediction model was carried out by integrating the introduced SHAP framework.

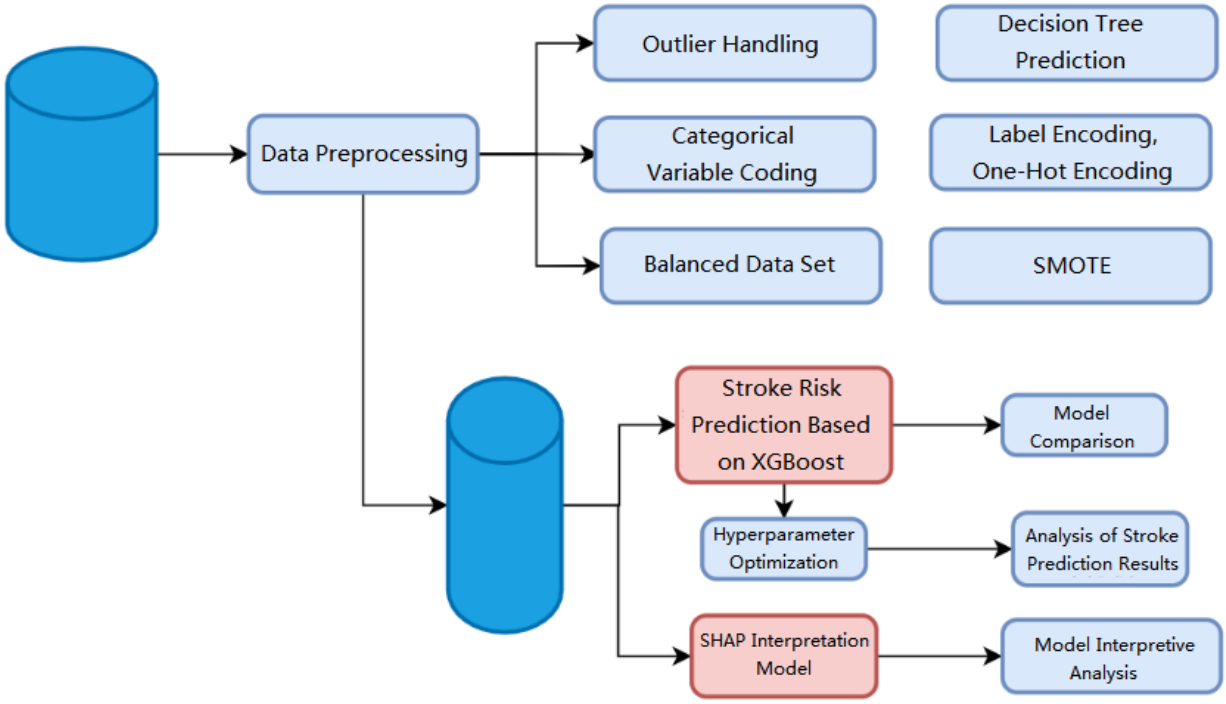

Figure 1. Process of constructing the interpretable stroke prediction model

3.1. Data source

Kaggle is a data science competition platform that involves data analysis, machine learning, deep learning, and many other fields. It contains massive data sets. The stroke prediction data set on this platform contains a total of 5,110 valid clinical samples, including 249 cases of stroke patients and 4,861 cases of non-stroke patients. The recorded variables include gender, age, history of hypertension, history of heart disease, marital status, occupation, type of residence, average blood glucose level, body mass index (BMI), and smoking history.

3.2. Data preprocessing

The categorical variables were encoded and converted into numerical variables. **Table 2** lists the relevant information of the variables after the data were encoded. For the 201 cases (4.0%) in which the BMI variable is missing, a decision tree was used to predict and fill, thus avoiding random noise. At the same

time, the garbage data whose age is not an integer were replaced and rounded by filling missing values. The ratio of the number of cases of stroke patients (249) to that of non-stroke patients (4,861) is close to 1:20, indicating that the categorical distribution of this data set is extremely unbalanced. Synthetic Minority Oversampling Technique (SMOTE) [9], an improvement over random sampling methods, can improve the class distribution of samples, while reducing the risk of overfitting. SMOTE was used on the stroke data set to alleviate the data imbalance problem, making the ratio of stroke patients to non-stroke patients close to 1:1.

Table 2. Relevant information of encoded variables

Variable	Type	Code/data value range
Demographics		
Gender	Unordered three categories	One-hot encoding
Age	Value type	[1,82]
Body mass index (kg/m^2)	Value type	[10.3,97.6]
Clinical history		
History of high blood pressure	Unordered binary classification	Yes = 1; No = 0
History of heart disease	Unordered binary classification	Yes = 1; No = 0
Average blood glucose level (mg/dL)	Continuous	[55.12,271.74]
Lifestyle		
Occupation	Unordered five categories	One-hot encoding
Smoking history	Unordered four categories	One-hot encoding
Type of residence	Unordered binary classification	City = 1; Country = 0
Marital status	Unordered binary classification	Married = 1; Single = 0

3.3. Prediction model

Ensemble learning completes learning tasks by constructing and combining multiple learners and achieves significantly superior generalization performance than a single learner [10].

The XGBoost algorithm is a boosting-based ensemble learning algorithm proposed by Chen *et al.* [11] in 2016. It is widely used in many machine learning and data mining problems owing to its fast execution speed and ability to guarantee model performance. In addition to the loss of prediction, the loss function of the XGBoost algorithm also adds a regularization part, which can prevent overfitting and has a stronger generalization ability. For each weak learner, such as the decision tree, parallel selection was made during the establishment process to find appropriate subtree node split features and eigenvalues, thereby improving operating efficiency. In this study, we established an XGBoost model to complete the stroke risk prediction analysis. In order to verify its performance, we compared it with several common machine learning algorithms, such as support vector machine (SVM), logistic regression (LR), decision tree (DT) and adaptive boosting (AdaBoost).

3.4. Parameter adjustment

In order to maximize the predictive ability of the model, it is necessary to adjust the parameters of the algorithm. Grid search is a commonly used hyperparameter adjustment technique. Following the arrangement and combination of parameters to generate a "grid," each set of parameters was used for algorithm training, and the prediction effect of the model was verified. Through grid search, the set of parameters with the best effect was selected. In order to avoid the influence of prediction randomness, the results of cross-validation are frequently used as the judging criteria in this process. Therefore, using

GridSearchCV in the machine learning library scikit-learn to perform hyperparameter optimization, the optimal parameter values of the XGBoost classifier were obtained as follows:

learning_rate = 0.01, max_depth = 6, n_estimators = 700, min_child_weight = 1, gamma = 0.5, colsample_bytree = 0.7, subsample = 0.8, reg_alpha = 0.05, reg_lambda = 0.5.

The experimental part of this study is in Kaggle Notebooks.

3.5. Model evaluation

The data set balanced by SMOTE were divided into two sets (training set and test set) in a 4:1 ratio. The training set was further divided into 10 parts, 9 parts of which were used for training the model each time, and the remaining part was used for evaluating the performance of the model. In this process, each sampling result was guaranteed to be different, and the performance index of the model under ten-fold cross-validation can be obtained by repeating 10 times. Although this method increases the amount of calculation, it greatly reduces the statistical uncertainty in model evaluation, which is conducive to the selection of models with excellent performance. Subsequently, we evaluated the model's predictions on the test set.
In this study, the four indicators accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the receiver operating characteristic (ROC) curve, namely AUC, were used to evaluate the predictive effect of the model on unseen data. The relevant indicators were calculated as follows:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

$$Spe = \frac{TN}{FP + TN} \tag{3}$$

where TP, FN, FP, and TN represent true positive cases, false negative cases, false positive cases, and true negative cases, respectively. The ROC curve can be used to evaluate the quality of the model; the abscissa is the false positive rate (FPR), and the ordinate is the true positive rate (TPR). The calculation formulas for both are as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

AUC corresponds to the area under the ROC curve, which considers the sorting quality of the sample prediction; the larger the value, the better the prediction performance of the model.

3.6. Model interpretation

In order to enhance their ability to solve problems, data scientists have proposed various algorithms to build complex models. However, the reasons behind the decisions made by these complex models are not known. Model interpretability aims to help interested parties understand the characteristics of the model and the reasons for decisions, thereby making the process more transparent. There is a large body of research work

devoted to exploring interpretability in the medical field. Gomolin *et al*. [12] pointed out that artificial intelligence, as a black box, is a "powerful limitation" of interpretability applications in dermatology, as it cannot be customized for evaluation by board-certified dermatologists to interpret clinical evidence. SHAP, as proposed by Lundberg *et al*. [13], is a unified framework for explaining forecasts. Its main idea comes from Shapley value in combinatorial game theory [14]. SHAP attributes the predicted value of the model to the Shapley value of each feature. Assuming that the *j* th feature of the *i* th sample is x_{ij}, the predicted value of the model for the *i* th sample is y_i, and the predicted mean of all samples is y_0, the formula is as follows:

$$y_i = y_0 + \sum_{j=1}^{m} f(x_{ij}) \tag{6}$$

where *m* is the number of features, with $f(x_{ij})$ representing x_{ij} the Shapley value. To determine the overall influence of a feature on the predicted value, the Shapley values of all samples corresponding to the feature can be averaged. In this study, we analyzed the results of the prediction model by drawing a SHAP-risk factor value scatter diagram.

4. Results

4.1. XGBoost model prediction

For the training set, a stroke risk prediction model was established with the XGBoost algorithm, and its performance was evaluated on the test set. In order to verify the superiority of the model's performance, the XGBoost model was compared with models constructed by four mainstream algorithms AdaBoost, DT, LR, and SVM. **Figure 2** is a boxplot of ten-fold cross-validation evaluation indicators based on accuracy on the training set. Among them, the average ten-fold cross-validation accuracy on the XGBoost model training set was the highest (97.2%), and the performance was stable (standard deviation [SD] = 0.0039); followed by the AdaBoost model in which the average ten-fold cross-validation accuracy on the training set was 96.2%; however, the fluctuation was larger (SD = 0.0068) compared to XGBoost; DT came in third in terms of performance, while LR and SVM had the worst performance. The results verified the performance of the XGBoost model.

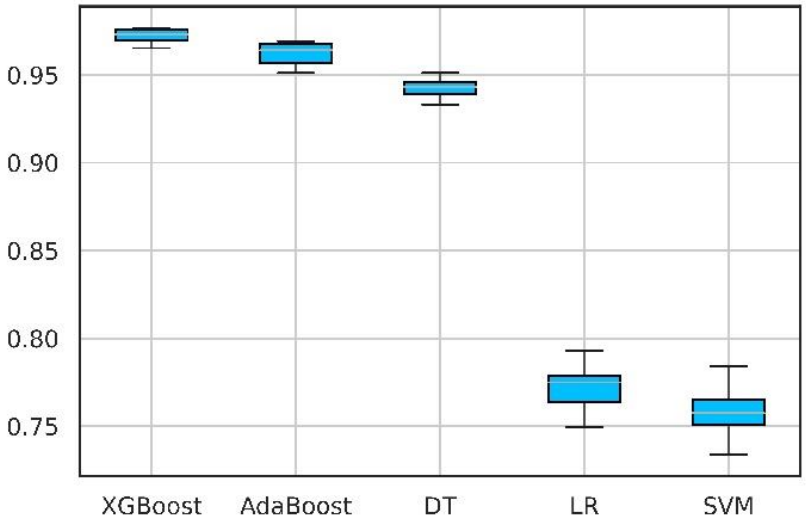

Figure 2. Box plot of cross-validation accuracy on training set

Figure 3 is the ROC curve that can reflect the prediction performance of different models. It can be seen that the ROC curve of the XGBoost model completely "wraps" the ROC curve of other mainstream models (SVM, LR, DT, and AdaBoost), suggesting that the overall prediction effect of the XGBoost model on the test set is the best.

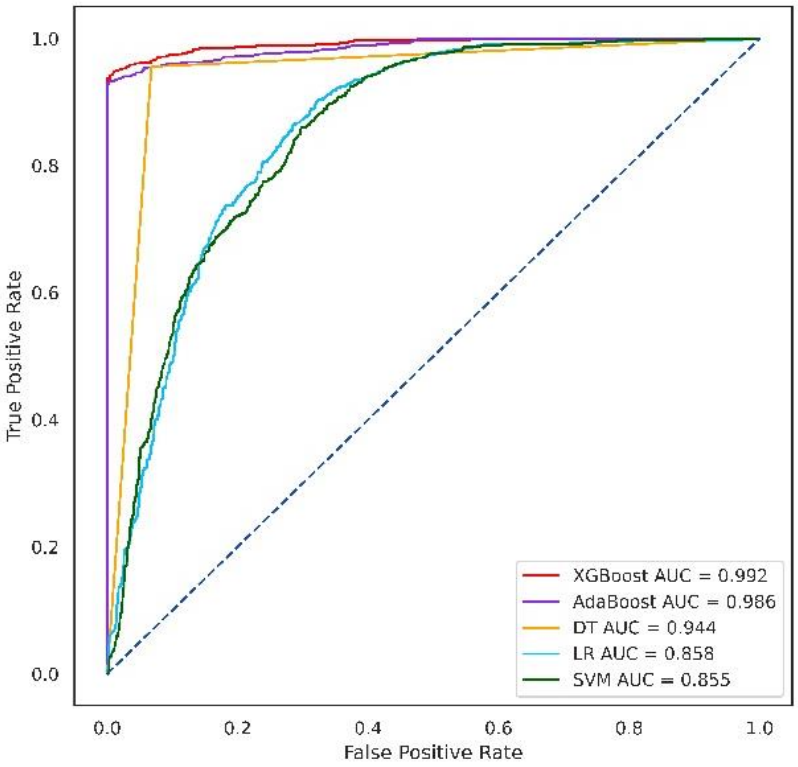

Figure 3. Receiver operating characteristic curves of different models on test set

Table 3 shows the results of performance comparison of different indicators among the five stroke risk prediction models on the test set. The overall prediction accuracy of the XGBoost model was 96.71%; its sensitivity, specificity, and AUC were 93.83%, 99.59%, and 99.19%, respectively (see the first row in **Table 4**). Among them, the overall accuracy and AUC were the highest among the five models, while its sensitivity and specificity were the second highest. In terms of overall performance, the XGBoost model performed the best. Comparing with existing research [6] (accuracy rate, 91.47%; sensitivity, 86.44%; specificity, 95.71%) and literature [15] (sensitivity, 89.7%; specificity, 72.7%; AUC, 87.8%), these indicators were higher, indicating that the model in our study can better predict the risk of stroke.

Table 3. Performance comparison of different models on test set

Model algorithm	Acc	Sen	Spe	AUC
XGBoost	**0.9671**	0.9383	0.9959	**0.9919**
AdaBoost	0.9635	0.9300	**0.9970**	0.9862
DT	0.9440	**0.9558**	0.9322	0.9440
LR	0.7820	0.8148	0.7492	0.8583
SVM	0.7722	0.8282	0.7163	0.8551

Abbreviations: AdaBoost, adaptive boosting; Acc, accuracy; AUC, area under the receiver operating characteristic curve; DT, decision tree; LR, logistic regression; Sen, sensitivity; Spe, specificity; SVM, support vector machine; XGBoost, extreme gradient boosting.

4.2. Model explanatory analysis based on SHAP

It can be seen that the stroke risk prediction model based on XGBoost has good predictive ability, but the contributing relationship between risk factors and risks can provide better support for auxiliary diagnosis analysis. In order to further analyze the positive/negative relationship between risk factors and output results, the integration of SHAP in this study can aid the explanatory analysis of the stroke risk prediction model based on XGBoost and demonstrate the SHAP-risk factor value distribution of the prediction model on the data set. The dot plot, as shown in **Figure 4**, directly reflects the impact of different risk factors of the sample and their values on the prediction results. Each point corresponds to the risk factor value of a sample. The color of the point is reminiscent of red when the value is larger. The abscissa represents the SHAP value of the point, *i.e.*, the impact on the output result. The vertical axis represents different risk factors; the higher the position of the risk factor, the greater the potential of the risk factor affecting the model output. It can be seen from the figure that among all risk factors, age affects the prediction results the most. This indicates that as age increases, the probability of stroke increases. The second is the type of residence; when other risk factors are kept constant, the risk of stroke is lower in rural residents. The likelihood of stroke is also higher with underlying heart disease or higher blood glucose level. All these are reflected in the high-risk groups of stroke in the stroke prevention report [1], which verifies the risk factor prediction analysis results of the model in this study. In addition, the figure also shows that non-smokers have a lower probability of stroke than those with a history of smoking, as mutually verified by Pandian *et al*. [2]. Thus, controlling tobacco should be made a priority as a primary stroke prevention strategy.

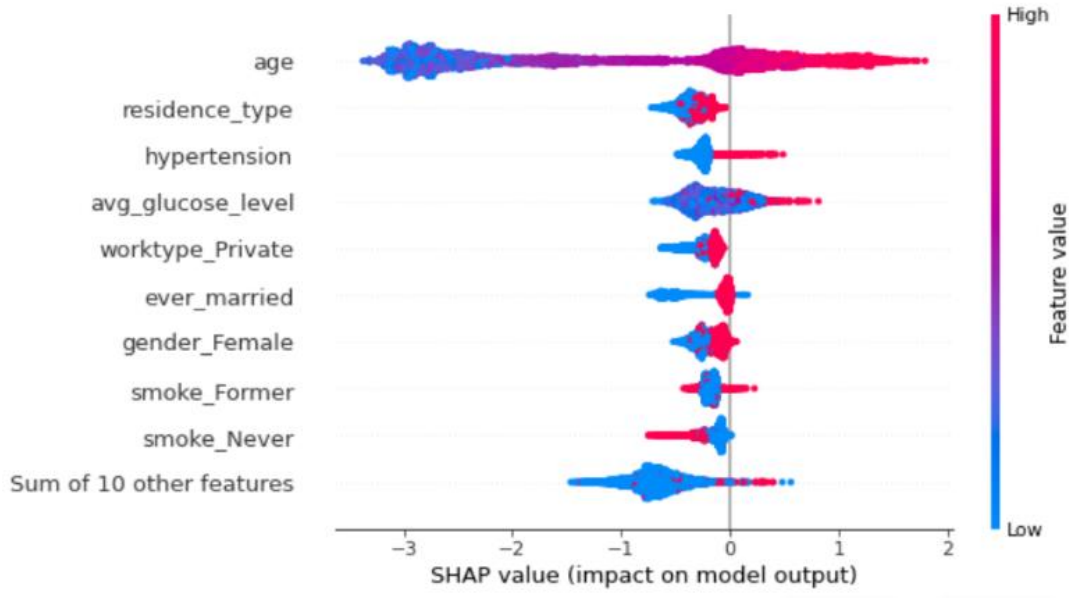

Figure 4. SHAP-risk factor distribution

Through the explanatory framework, we can further analyze how a single risk factor affects the model output and the interaction between different risk factors. Taking age, which has the most significant impact on the model output, as an example, a SHAP-risk factor dependency diagram was drawn, as shown in **Figure 5**. It can be seen that before the age of 40, aging hardly affects the risk of stroke, but after that, the risk of stroke increases linearly with age. High blood glucose level accounts for less in the lower age group, reducing the SHAP value at this time, but the opposite is true in the higher age group.

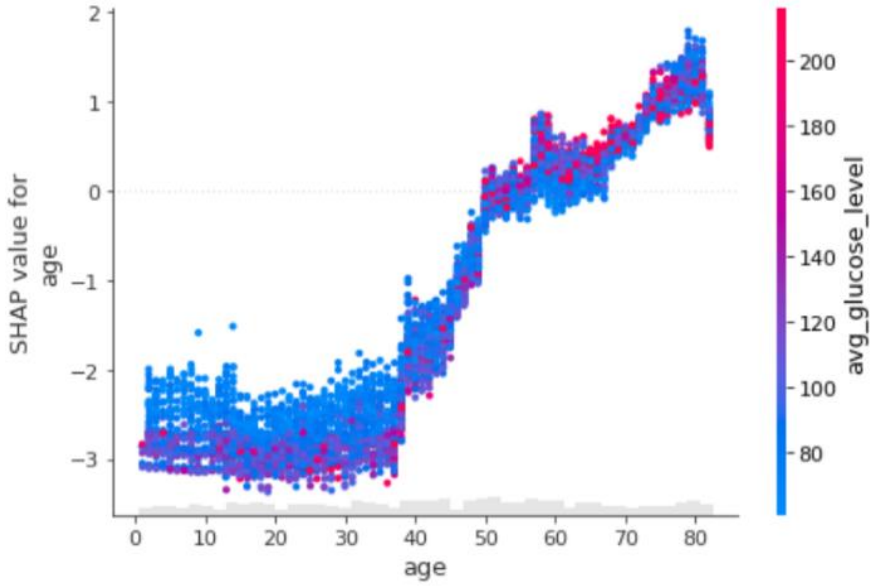

Figure 5. SHAP-risk factor (age) dependency diagram

Through the explanatory framework, we can also analyze the prediction results of the model for a single sample. **Figures 6** and **7** are interpretations for the prediction of stroke patients and non-stroke patients, respectively. The red part represents the contribution of risk factor values to the prediction of stroke, while the blue part represents the contribution of risk factor values to the prediction of non-stroke. **Figure 6** shows that high body mass index, high blood glucose level, old age, and heart disease are the main factors contributing to the prediction of stroke. **Figure 7** shows that the reasons for the prediction of non-stroke include the absence of heart disease, freelancing, and no history of hypertension.

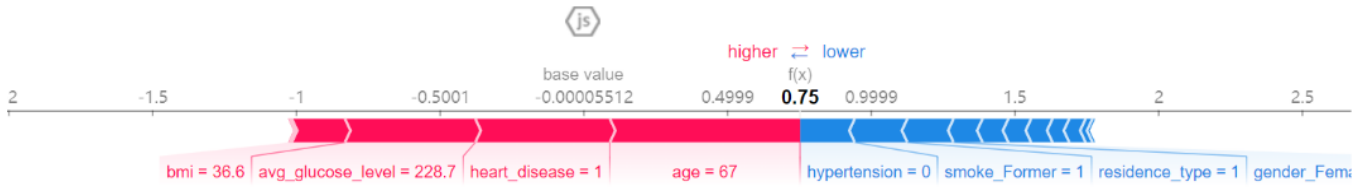

Figure 6. Example of SHAP interpretation predicted as stroke

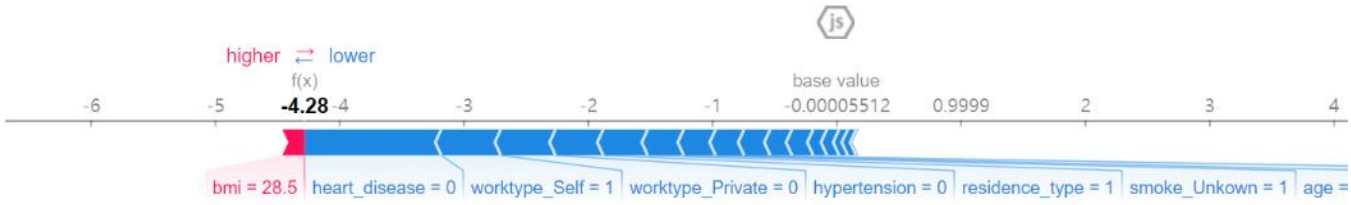

Figure 7. Example of SHAP interpretation predicted as non-stroke

Figure 8 presents the average SHAP values of different risk factors and their sorting results. For the XGBoost model, age is the most important factor affecting individuals suffering from stroke, followed by the type of residence, history of hypertension, and average blood glucose level. These important risk factors observed based on SHAP are consistent with the results demonstrated by existing studies [16]. It is worth noting that marital status is also an important factor in the prediction model of this study, thus suggesting that this component should be taken into consideration as a risk factor in research.

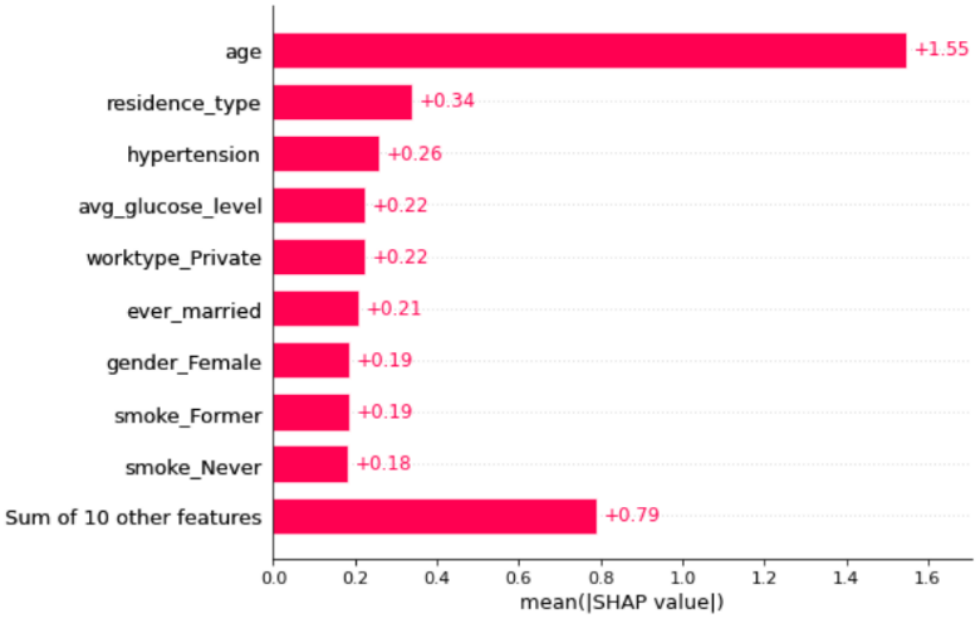

Figure 8. Ranking of risk factor based on SHAP

5. Conclusion

In recent years, stroke has become the second leading cause of death and disability worldwide [17]. In order to screen out risk factors more scientifically and actualize risk prediction and early warning to carry out early prevention work, we propose an integrated model that uses the XGBoost algorithm to establish a risk prediction model and the SHAP explanatory framework for model prediction analysis.

(i) Through outlier processing, feature variable coding, data category balance, and other data preprocessing and grid search parameter optimization, we established the XGBoost risk prediction model and compared it with four mainstream algorithm models. The performance of the model in predicting stroke risk was verified in this study.

(ii) By constructing an explanatory analysis model based on the integration of the XGBoost risk prediction model with SHAP, an explanatory analysis of the contribution of risk factors to risk prediction was performed. It can further assist medical workers in further diagnostic analysis. The results of this study showed that age, type of residence, history of hypertension, *etc*., are important factors affecting the incidence of stroke.

In future practice, we shall incorporate more risk factors to explore the prediction of stroke risk and the contribution of risk factors to stroke risk, as well as use clinical data to conduct more in-depth model research to better assist medical workers in realizing early warning and prevention of stroke risk.

Funding

This work was supported the National Natural Science Foundation Project (Grant No. 61863027), the Special Research Project on High Quality Development of Innovation and Entrepreneurship Education of the Chinese Society of Higher Education (Grant No. 21CXD01), and the Key R&D Plan of Jiangxi Province (Grant No. 20202BBGL73057).

Disclosure statement

The authors declare no conflict of interest.

References

[1] The Writing Group of "China Stroke Prevention Report", 2022, Summary of "China Stroke Prevention Report 2020". Chinese Journal of Cerebrovascular Diseases, 19(2): 136–144.

[2] Pandian JD, Gall SL, Kate MP, et al., 2018, Prevention of Stroke: A Global Perspective. The Lancet, 392(10154): 1269–1278.

[3] Wang Y, 2019, Research on Influencing Factors and Risk Prediction Model of Stroke Based on Big Data, thesis, Guangdong University of Technology.

[4] Barra S, Almeida I, Caetano F, et al., 2013, Stroke Prediction with an Adjusted R-CHA2DS2VASc Score in a Cohort of Patients with a Myocardial Infarction. Thrombosis Research, 132(2): 293–299.

[5] Vartiainen E, Laatikainen T, Peltonen M, et al., 2016, Predicting Coronary Heart Disease and Stroke: The FINRISK Calculator. Global Heart, 11(2): 213–216.

[6] Hou Y, Zhang C, Su Y, 2019, Risk Prediction of Ischemic Stroke Based on Support Vector Machine. Modern Preventive Medicine, 46(15): 2692–2695 + 2700.

[7] Luo Y, Shao Y, Chen D, 2021, Prediction of Annual Stroke Risk of Ischemic Stroke Based on BiLSTM-Attention Model. Journal of Donghua University (Natural Science Edition), 47(4): 62–68.

[8] Chen J, Chen Y, Li J, et al., 2022, Stroke Risk Prediction with Hybrid Deep Transfer Learning Framework. IEEE Journal of Biomedical and Health Informatics, 26(1): 411–422.

[9] Chawla NV, Bowyer KW, Hall LO, et al., 2002, SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16: 321–357.

[10] Zhou Z-H, 2021, Ensemble Learning, in Machine Learning, Springer, Singapore, 181–210.

[11] Chen T, Guestrin C, 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13–17, 2016: XGBoost: A Scalable Tree Boosting System, ACM, San Francisco California USA, 785–794.

[12] Gomolin A, Netchiporouk E, Gniadecki R, et al., 2020, Artificial Intelligence Applications in Dermatology: Where Do We Stand?. Frontiers in Medicine, 7: 100.

[13] Lundberg SM, Lee S-I, 2017, A Unified Approach to Interpreting Model Predictions, in Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 4765–4774.

[14] Shapley LS, 1952, A Value for N-Person Games, RAND Corporation, Santa Monica, CA.

[15] Li M, Wang C, Xia B, et al., 2017, Stroke Risk Prediction Model for Health Management Population. Journal of Shandong University (Medical Science), 55(6): 93–97 + 103.

[16] Boehme AK, Esenwa C, Elkind MSV, 2017, Stroke Risk Factors, Genetics, and Prevention. Circulation Research, 120(3): 472–495.

[17] Murphy SJX, Werring DJ, 2020, Stroke: Causes and Clinical Features. Medicine, 48(9): 561–566.