Development and Validation of Interpretable Machine Learning for Stroke Occurrence in Older, Community Chinese Dwellers

Yafei Wu
The State Key Laboratory of Molecular Vaccine and Molecular Diagnostics, School of Public Health, Xiamen University
Zhongquan Jiang
The State Key Laboratory of Molecular Vaccine and Molecular Diagnostics, School of Public Health, Xiamen University
Shaowu Lin
The State Key Laboratory of Molecular Vaccine and Molecular Diagnostics, School of Public Health, Xiamen University
Ya Fang (fangya@xmu.edu.cn)
The State Key Laboratory of Molecular Vaccine and Molecular Diagnostics, School of Public Health, Xiamen University

Research Article

Keywords: stroke, machine learning, prediction, older adults

Posted Date: June 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-604690/v1

Abstract

Background: Prediction of stroke based on individuals' risk factors, especially for a first stroke event, is of great significance for primary prevention of high-risk populations. Our study aimed to investigate the applicability of interpretable machine learning for predicting a 2-year stroke occurrence in older adults compared with logistic regression.

Methods: A total of 5960 participants consecutively surveyed from July 2011 to August 2013 in the China Health and Retirement Longitudinal Study were included for analysis. We constructed a traditional logistic regression (LR) and two machine learning methods, namely random forest (RF) and extreme gradient boosting (XGBoost), to distinguish stroke occurrence versus non-stroke occurrence using data on demographics, lifestyle, disease history, and clinical variables. Grid search and 10-fold cross validation were used to tune the hyperparameters. Model performance was assessed by discrimination, calibration, decision curve and predictiveness curve analysis.

Results: Among the 5960 participants, 131 (2.20%) of them developed stroke after an average of 2-year follow-up. Our prediction models distinguished stroke occurrence versus non-stroke occurrence with excellent performance. The AUCs of machine learning methods (RF, 0.823[95% CI, 0.759-0.886]; XGBoost, 0.808[95% CI, 0.730-0.886]) were significantly higher than LR (0.718[95% CI, 0.649, 0.787], $p<0.05$). No significant difference was observed between RF and XGBoost ($p>0.05$). All prediction models had good calibration results, and the brier score were 0.022 (95% CI, 0.015-0.028) in LR, 0.019 (95% CI, 0.014-0.025) in RF, and 0.020 (95% CI, 0.015-0.026) in XGBoost. XGBoost had much higher net benefits within a wider threshold range in terms of decision curve analysis, and more capable of recognizing high risk individuals in terms of predictiveness curve analysis. A total of eight predictors including gender, waist-to-height ratio, dyslipidemia, glycated hemoglobin, white blood cell count, blood glucose, triglycerides, and low-density lipoprotein cholesterol ranked top 5 in three prediction models.

Conclusions: Machine learning methods, especially for XGBoost, had the potential to predict stroke occurrence compared with traditional logistic regression in the older adults.

Background

Stroke is the leading cause of death and disability worldwide [1], with a substantial treatment and prognostic care costs [2]. As the disease spectrum has transited from infectious and malnutrition diseases to noninfectious chronic diseases (NCDs), together with the frequent exposure to unhealthy lifestyle and environmental pollution, the global burden of stroke will continue to rise [3]. It is estimated that there will be approximately 200 million stroke patients worldwide in 2050, thereafter, 30 million new cases and 12 million deaths every year without effective prevention measures [4]. Stroke has been identified as one of the prioritized diseases in the World Health Organization and the United Nations on NCDs [5]. With the rapid aging of population, stroke has also become a huge challenge in China, with an increase of 2 million new cases every year [6]. Therefore, development of effective stroke prediction models for guiding early identification of high-risk populations is an urgent task.

Stroke-related predictions generally include four aspects: stroke prevention or risk factor identification, stroke diagnosis, stroke treatment and stroke prognosis [7]. So far, most studies focused on stroke diagnosis [7-11], which required complex medical data, such as physical, neurological, and brain imaging examination (CT or MRI) to exclude other diseases (stroke mimics), and to recognize its type, location and severity [12]. Actually, the imaging data are usually obtained when stroke has already occurred, and imaging examinations are expensive, which is undoubtedly unsuitable for early screening and reduction of costs [13]. Additionally, the prognosis of stroke is often poor since there is no effective treatments. Thus, early identification of high-risk individuals and personalized interventions are the most cost-effective way [14]. Fortunately, with the increase in population-based cohort studies, it's easy to obtain individuals' macro data (epidemiological data) through questionnaires as well as the micro data, such as blood biomarkers. A full utilization of the comprehensive data derived from population-based cohort would be helpful for early identification of high-risk populations of stroke. Many previous studies, such as the UKPDS calculator [15], PROCAM calculator [16], SCORE risk table [17], ASCVD calculator [18], QRISK calculator [19], etc., predicted the

risk of cardiovascular disease using demographics, biomarkers, and clinical variables. While only few studies [20-22] focused on stroke prediction, especially for adults aged 60 and above.

Regression methods, such as Cox regression and logistic regression with simplicity and interpretability, were the commonly used prediction methods in previous studies. Traditional regression methods mainly deal with low-order interaction, such as first-order interaction effects [20, 21, 23], making it difficult for analyzing high-order nonlinear relationships. Especially when number of predictors or the explanatory ability of predictors is limited, the complicated relationships between predictors and outcomes may not be captured [24]. Machine learning (ML), which is a set of computational methods that can discover complex nonlinear relationships between inputs and outputs, has been widely used in the field of disease prediction and health research [25-27]. Among ML methods, ensemble learning was a widely used method with excellent performance [7, 28], which makes predictions through integrating the results of multiple weak classifiers. Logistic regression (LR), a representation of regression methods, was often used as a reference model to compare with other ML methods.

Here, we developed two interpretable ensemble learning methods, namely random forest (RF) and extreme gradient boosting (XGboost), to predict 2-year stroke outcome (stroke occurrence versus non-stroke occurrence) in the elderly aged 60 and over compared with logistic regression based on demographics, lifestyle, disease history and blood biomarker data. Specifically, the predictive performance was assessed by discrimination, calibration, decision curve and predictiveness curve. Besides, the interpretable machine learning techniques were used to understand the predictors of black-box ML methods towards clinical practice.

Methods

Data source

This study retrospectively collected data from the China Health and Retirement Longitudinal Study (CHARLS) from July 2011 to August 2013 (*http://charls.pku.edu.cn/index/zh-cn.html*). The detailed information of CHARLS were described elsewhere [29]. A series of health data in 2011 was used as potential predictors, and the self-reported physician diagnosis stroke status was collected in the 2013 follow up wave and was used as a binary outcomes (stroke occurrence versus non-stroke occurrence). Participants were included in study, if (1) aged 60 years or above in baseline, (2) without stroke in baseline. Participants with missing value for stroke status were further excluded. Finally, 5960 participants were eligible for analysis. Among them, 131 participants reported having a stroke after a 2-year follow-up.

Data preprocessing

Data on 16 variables were collected before constructing the prediction models, including demographics (age, gender, and waist-to-height ratio); lifestyle (smoking, drinking); disease history (hypertension, diabetes, dyslipidemia, and heart disease); clinical variables (high sensitivity C-reactive protein, white blood cell, glucose, glycated hemoglobin, low density lipoprotein cholesterol, triglycerides, and cystatin C). Waist-to-height ratio (WHtR) was calculated by dividing waist (cm) by height (cm). Smoking and drinking were converted into binary variables (1 for yes, 0 for no). Disease history was collected from the self-report-based physician's diagnosis, and was also treated as binary variables (1 for yes, 0 for no). For missing values, imputation was performed with two strategies. In logistic regression, continuous variables were imputed with median, and categorical variables were imputed with mode. In random forest and XGBoost, the imputation was processed by the algorithm itself, where continuous predictors were imputed by the weighted average of non-missing values and categorical variables were imputed with the class with largest average proximity. Additionally, we observed that data was quite imbalanced in study, i.e. the ratio between non-stroke and stroke population (about 44) was far from 1. Therefore, the Synthetic Minority Oversampling Technique (SMOTE), which could analyze the minority samples and synthesize new samples according to the minority samples [30, 31], were used for data balancing.

Feature selection with Boruta

Boruta algorithm, a commonly used feature selection method, was further used to select a few more relevant predictors for constructing prediction models. Boruta is a wrapper method built with RF classifier. It is an extension from the thought of Stoppiglia, Dreyfus, Dubois and Oussar [32], and can determine the importance of variables by comparing the correlation between real features and shadow features. Traditional feature selection algorithms are more likely to leave out some relevant features in the process of minimizing errors because of its minimum optimal criteria. While Boruta can find all features with a full correlation strategy, that is, even predictors weakly related to outcome were preserved [33]. The main steps of Boruta algorithm were as follows: (1) based on each real feature R, randomly shuffle the order and construct a new feature (shadow feature S), and connect the shadow feature behind the real feature to obtain a new feature matrix $N = [R, S]$. (2) take the new feature matrix N as input, train the data with RF model, and output the variable importance. (3) select the real feature with variable importance higher than shadow feature (Z score of the real feature is larger than the maximum Z score of the shadow feature) in each iteration, and removed the unimportant real features. (4) the algorithm stops when all features are compared or reaches the maximum number of iterations.

Marginal variables that are on the edge of acceptance and rejection may exist when implementing Boruta algorithm. In order to quantitatively evaluate the effects of marginal variables on predictions, we constructed a reference model with the accepted variables, then a new prediction model incorporating both the accepted variables and marginal variables were constructed, and net reclassification improvement (NRI) [34] and integrated discrimination improvement (IDI) [35] were used to assessed the contribution of marginal variables to predictions. A positive value of NRI or IDI indicates an improvement of performance.

Prediction models

In this study, logistic regression was used as the reference model for comparisons with ML methods. The binary LR had only two possible values for outcome (stroke occurrence versus non-stroke occurrence). RF and XGBoost were selected as the representative interpretable ensemble learning models. RF, one of the commonly used bagging methods, was proposed by Leo Breiman [36], which could generate multiple decision tree classifiers in a parallel manner, and eventually achieved classification with minority voting or numerical prediction through averaging. XGBoost, proposed by Chen et al. [37], is another kind of ensemble learning known as boosting strategy. The construction of next weak classifier in boosting depends on its previous classifier. XGBoost fits the predicted residuals with multiple weak classifiers, and finally synthesizes all weak learners to obtain a powerful learner.

Model derivation and internal validation

Data divided with a ratio of 7:3 were used for model derivation and internal validation, respectively. In derivation stage, 10-fold cross validation and grid search were used to tune hyperparameters. In validation stage, accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were used to assess models' discrimination. Bootstrap method was used to calculate the 95% confidence interval (CI) of AUC, and AUC between all models were also compared. Furthermore, the calibration of each prediction model was assessed by brier score. We also performed decision curve analysis (DCA) [38] and predictiveness curve analysis [39] towards clinical usefulness, which could provide reference for selection of optimal model. DCA, proposed by Andrew J. Vickers et al. in 2006, is a simple method for evaluating prediction models, which considers both accuracy and clinical usefulness. Predictiveness curve describes the cumulative proportion towards absolute risk. The lower the proportion of the intermediate risk (between high-risk and low-risk), the better the model is able to distinguish the high- and low-risk populations. Additionally, all predictors were ranked according to its importance in each model, and we further used SHAP (Shapley Additive exPlanation) to understand the interpretability of machine learning models. SHAP is derived from game theory based on Shapley value, which can not only show the variable importance, but also determine the direction of effects [40]. The whole process of derivation and validation was shown in Figure 1.

Statistical analysis

Continuous variables were presented as mean±standard deviation (normal distribution), and as median with interquartile range (IQR, skewed distribution). Categorical variables were presented as percentages. All analyses were performed with R3.6.0. A

two-sided *p*-value of <0.05 was considered statistically significant.

Results

Baseline characteristics

The 2-year prevalence of stroke in the whole population was 2.20%, with women experiencing relatively higher than men (2.38% vs 2.02%). In terms of incidence intensity, an average of 1.10 strokes per 100 person-years was observed in 2-year follow-up, and women was also higher than men (1.19 / 100 person year versus 1.01 / 100 person year).

Comparisons of baseline characteristics are shown in Table 1. Briefly, the average age of participants was 66.75 years old, and almost half of them were females. The proportion of smoking and drinking populations was relatively high (41.49% and 30.49%, respectively). For chronic diseases, hypertension was quite common for older adults (30.74%), followed by heart disease (14.88%), dyslipidemia (9.49%), and diabetes (6.51%). Stroke patients were more likely to be older men with smoking habits and chronic diseases. However, the proportion of drinking alcohol was much lower in stroke populations. Other characteristics were balanced between two groups.

Selection of predictors using Boruta

The results of feature selection based on Boruta are shown in Figure 2. The three blue features represented the maximum *Z* score, average *Z* score, and minimum *Z* score of shadow feature. Features in green indicated accepted features (GLU, WBC, HbA1c, TG, CysC, hsCRP, LDLC, Age, WHtR, Dyslipidemia, Gender, and Smoking), features in red represented rejected features (heart_disease and drinking), and features in yellow represented marginal features (hypertension and diabetes). Furthermore, NRI and IDI were calculated for assessing the contributions of marginal features (Table2). The results showed that the predictive performance did not significantly improve ($p>0.05$) after adding two marginal features. Meanwhile, adding more predictors into prediction model would make it complicated, therefore, the 12 accepted features were finally selected for model derivation.

Comparisons between models in internal validation

Comparisons of discrimination and calibration

For logistic regression, the default hyperparameter was used. For RF, *mtry* and *ntree* were tuned. For XGBoost, a series of hyperparameters including learning rate (*eta*), proportion of features used by each split node in each layer of the tree (*colsample_bylevel*), sampling ratio (*subsample*), the max depth of tree (*max_depth*), the minimum sample weight required by the leaf node (*min_child_weigt*), the minimum drop of loss function required for node splitting (*gamma*) were tuned. Finally, a combination of *mtry* = 5, *ntree* = 500 were selected as the optimal parameters for RF, for XGBoost, the best combination was as follows: *eta* = 0.05, *colsample_bylevel* = 2/3, *subsample* = 0.75, *max_depth* = 5, *min_child_weigt* = 2, *gamma* = 0.1. The predictive performance is shown in Table 3. We found that accuracy of the three models were close to each other (LR: 0.738, RF: 0.743, XGBoost: 0.740). In terms of sensitivity, XGBoost (0.700) ranked first, followed by RF (0.750) and LR (0.525). For specificity, LR and RF were similar (0.726), both were higher than XGBoost (0.724). The AUCs of machine learning methods (RF, 0.823[95% CI, 0.759-0.886]; XGBoost0.808[95% CI, 0.730-0.886]) were higher than LR (0.718[95% CI, 0.649-0.787], $p<0.05$). But there was no significant difference between RF and XGBoost ($p>0.05$).

As for calibration, both LR and machine learning methods generally calibrated well (Table 3). The brier score of LR, RF, and XGBoost were 0.022 (95% CI, 0.015-0.028), 0.019 (95% CI, 0.014-0.025), and 0.020 (95% CI, 0.015-0.026), respectively.

Comparisons of clinical usefulness

For DCA analysis (Figure 3-A), the horizontal dashed line was based on the assumption that all participants were free of stroke, and its net benefit was 0. Conversely, the oblique dashed line showed net benefits at different thresholds with the assumption that all participants were stroke patients. We found that net benefits of machine learning methods were generally much higher

than logistic regression, especially when threshold exceeded 0.08%, the benefits of logistic regression transited to be negative, while machine learning methods still had higher net benefits within a wider threshold range. Figure 3-B showed the predictiveness curves of three prediction models. The horizontal dashed line represented the prevalence of stroke (4.94%) in the elderly aged 60 and above in China. If the value of 4.94% was used as a threshold for identifying high and low risk population (binary classification), the proportions of high-risk groups were 6.484%, 13.583%, and 16.378% in LR, RF, and XGBoost, respectively. If a 2%-5% range was used as the intermediate risk (three categories), the proportions of the intermediate risk population were 32.700%, 20.291%, and 13.751% in LR, RF, and XGBoost, respectively.

In summary, XGBoost had relatively higher sensitivity, better calibration, and higher benefits within a larger threshold range.

Variable importance for prediction models

Results of variable importance are shown in Table 4. For LR, the top 5 predictors for predicting 2-year stroke occurrence were dyslipidemia, HbA1c, gender, WHtR, and WBC. Among them, all predictors except for HbA1c (OR 0.648) would significantly increase the risk of stroke (OR ≥1.28). For RF, the top 5 predictors were HbA1c, GLU, WBC, TG, and LDLC, and their proportion of importance accounted for 16.32%, 14.71%, 13.61%, 10.11%, and 9.71%, respectively. For XGBoost, the top 5 were GLU, HbA1c, WBC, LDLC, and WHtR, with their importance accounting for 15.02%, 13.31%, 11.11%, 10.41%, and 9.61%, respectively. HbA1c and WBC were the common predictors in the top 5 for all three models.

From the perspective of the consistency of predictors in three models, the proportion of complete consistency in three models was 0, and similar results were observed between LR and RF. A consistency of 8.33% was observed between LR and XGBoost. Surprisingly, the proportion of consistency reached 41.67% between the two machine learning methods. Especially when the ranking difference was allowed within 3, the consistency reached 91.67%.

We further used the SHAP to understand the interpretability of XGBoost (Figure 4). The results showed that the impact of WHtR, HbA1c, CysC, LDLC, and hsCRP on stroke was quite complicated, that is, the stroke risk showed positive relationship with the above predictors within a certain range, while the stroke risk showed negative relationship beyond that range. The impact of other predictors on stroke was mainly one-way, that is, as the exposure level increased, the risk of stroke increased. This also proved that ML methods were able to capture the complex nonlinear relationships contained in data.

Discussion

Elderly people are vulnerable to cardiovascular disease, such as stroke. Early risk identification and effective prevention were quite necessary for high-risk populations. Our study predicted the 2-year stroke occurrence with comprehensive data obtained from a population-based cohort. The results showed that machine learning could effectively distinguish stroke occurrence versus non-stroke occurrence in elderly individuals.

Sufficient data preprocessing, such as imputation, feature selection, data balancing, etc. was necessary before constructing predictive models [41, 42]. In a recent review study, the author pointed out that there were still many studies that had not addressed the above issues well [28]. Let’s take data balancing for example, the ratio of non-stroke and stroke patients was about 44 in the original data set, indicating a quite imbalanced distribution of outcome. ML methods would classify most of the participants into non-strokes for a high accuracy if trained directly in the imbalanced data set. In fact, we were cheated by them because of much lower performance in sensitivity and AUC except for accuracy and specificity. What’s worse, the model would perform badly when it was applied in different populations. Therefore, we calculated the ratio of specificity to sensitivity (sep/sen) to monitor the influence of SMOTE algorithm on performance. If sep/sen was close to 1, it meant that prediction model achieved a balance between sensitivity and specificity. Our results just proved that the sep/sen of LR, RF and XGBoost models were 1.79, 1.307, and 1.487, respectively without balancing , and were 1.383, 0.968, 1.034 after SMOTE processing. In addition, let’s have a look at feature selection. After implementing Boruta algorithm, more concise and more predictive predictors were obtained, which was crucial for constructing powerful ML models [43].

The traditional statistical method (LR) and two machine learning methods (RF, XGBoost) showed good performance in this study. The AUC of LR, RF and XGBoost were 0.718, 0.823, and 0.808, respectively. Marnie E. Rice et al. mentioned that AUC could be converted into the effect size, such as Cohen's *d* and the point-biserial correlation coefficient (r_{pb}) [44]. Cohen'*d* values for LR, RF and XGBoost in our study were 0.806-0.820, 1.30-1.33, 1.22-1.24, respectively, and r_{pb} were 0.374-0.379, 0.545-0.554, 0.520-0.528, respectively. According to the criterion of effect intensity towards Cohen'*d*, our predictive models were all equivalent to high effect levels. According to the criterion of effect intensity towards r_{pb}, the traditional LR represented the medium effect, and the two machine learning models still represented high effect level. Furthermore, we found that ML methods performed better than traditional regression models, as demonstrated in lots of previous studies [45-47]. While there were also some studies showed that performance of ML and regression models was comparable [48]. Some possible explanation might be as follows: ML is excellent in processing big data, so it may not find the complex rules when data is limited; besides, the selection of optimal predictors was also a tough task. Margaret S. Pepe et al. pointed out that influencing factors and predictors often had great contradictions even in the same research with same data. In other words, a factor may be closely related to disease, but it may contribute less in prediction research [49], which called attention to us that there was no absolutely superior model, so was ML methods. In practice, we should select the most suitable model in specific scenarios.

We found that a total of 8 variables ranked top 5 in three prediction models, namely gender, WHtR, dyslipidemia, HbA1c, WBC, GLU, TG, and LDLC, were significant for predicting 2-year stroke in older adult, which has important implications for stroke prevention. The aim of stroke prevention is to reduce the risk of developing a first stroke event through targeted modification of single or multiple modifiable risk factors at the population or individual level. Specifically, there may be two broad levels for stroke prevention: (1) primordial prevention was suitable for implementation on the population level, and targeted measures, such as healthy diet, physical exercise, weight control, and healthy lifestyle, were encouraged for prevention; (2) primary prevention was conducted on the individual level with a more personalized prevention strategies, such as changing the specific unhealthy lifestyle as well as identifying and treating chronic disease (ie, dyslipidemia) [50]. Among the 12 predictors in our study, all the predictors except age and gender were modifiable factors, which suggested that much work related to the modifiable risk factors needed to be done in future.

Evaluation of the clinical usefulness of predictive models could guide clinical practice. We performed decision curve and predictiveness curve analyses in this study, and the results showed that XGBoost was relatively superior in terms of clinical benefits and ability for distinguishing stroke patients. The practical significance of this study mainly reflected in the primary screening of high-risk individuals. Specifically, the ML models could be used to assess individuals' stroke risks with epidemiological data obtained from questionnaires as well as biomarkers through routine blood sampling. A healthy lifestyle and regularly health check in primary health center were recommended for individuals with low stroke risks. While for high-risk individuals, it was recommended to go to a high-level medical institution for examination to further determine whether there was a stroke.

Our study may have some potential advantages. First, we tried to construct predictive models with the comprehensive data derived from a population-based cohort, which was easy to obtain and low in cost, so it might be suitable for the primary screening of high-risk populations. Second, we constructed and evaluated prediction model for elderly people aged 60 years and older, for whom the stroke risk were much higher. Third, we performed relatively complete data preprocessing to ensure the quality of predictive models, providing a solid foundation for model construction. Fourth, we made a comprehensive assessment of prediction models, including discrimination, calibration, and the clinical usefulness analysis (decision curve and predictiveness curve), and SHAP for an in-depth discussion on the interpretability of ML models. Finally, we followed the standard reporting process of prediction model as described in TRIPOD [51].

However, there were still some limitations in our study. First of all, limited by data availability, the participants included in our study was not large enough, while ML is more powerful in processing big data, thus, the complex rules might not be discovered with limited data. In addition, we only evaluated the generalization ability of predictive models with internal validation, and an external validation in another populations is needed in future studies.

Conclusions

Based on the epidemiological and clinical data derived from population-based cohort, machine learning methods could effectively predict stroke occurrence in the adults aged 60 years and older. With a comprehensive consideration of ability to distinguish stroke occurrence and clinical benefits, XGBoost could be used for primary screening of high-risk individuals in community.

Abbreviations

noninfectious chronic diseases (NCDs); logistic regression (LR); random forest (RF); extreme gradient boosting (XGBoost); interquartile range (IQR); synthetic minority oversampling technique (SMOTE); integrated discrimination improvement (IDI); net reclassification improvement (NRI); confidence interval (CI); decision curve analysis (DCA); area under the receiver operating characteristic curve (AUC); machine learning (ML).

Declarations

Ethics approval and consent to participate

Data used in our study was approved by the biomedical ethics committee of Peking University, and all participants provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

Data used in this study could be obtained from the following link *http://charls.pku.edu.cn/index/zh-cn.html*.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the National Natural Science Foundation of China (No. 81973144), and the Field Investigation Foundation of Xiamen University (No. 2019GF032).

Authors' contributions

Ya Fang and Yafei Wu designed the research; Zhongquan Jiang and Yafei Wu analyzed the data; Yafei Wu drafted the initial manuscript. Ya Fang, Shaowu lin, and Zhongquan Jiang critically revised the draft manuscript. Ya Fang and Yafei Wu obtained funding. All authors read and approved the final manuscript.

Acknowledgments

We thank the China Health and Retirement Longitudinal Study (CHARLS) for providing us with the data.

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC, Beaton AZ, Benjamin EJ, Benziger CP *et al*: Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J AM COLL CARDIOL* 2020.

2. Rajsic S, Gothe H, Borba HH, Sroczynski G, Vujicic J, Toell T, Siebert U: Economic burden of stroke: a systematic review on post-stroke care. *EUR J HEALTH ECON* 2019, 20(1):107-134.

3. Johnson CO, Nguyen M, Roth GA, Nichols E, Alam T: Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *LANCET NEUROL* 2019, 18(5):439-458.

4. Brainin M, Feigin VL, Norrving B, Martins S, Hankey GJ, Hachinski V: Global prevention of stroke and dementia: the WSO Declaration. *LANCET NEUROL* 2020, 19(6):487-488.

5. Feigin VL, Norrving B, Mensah GA: Global Burden of Stroke. *CIRC RES* 2017, 120(3):439-448.

6. Wu S, Wu B, Liu M, Chen Z, Wang W, Anderson CS, Sandercock P, Wang Y, Huang Y, Cui L *et al*: Stroke in China: advances and challenges in epidemiology, prevention, and management. *LANCET NEUROL* 2019, 18(4):394-405.

7. Sirsat MS, Ferme E, Camara J: Machine Learning for Brain Stroke: A Review. *J Stroke Cerebrovasc Dis* 2020, 29(10):105162.

8. Karthik R, Gupta U, Jha A, Rajalakshmi R, Menaka R: A deep supervised approach for ischemic lesion segmentation from multimodal MRI using Fully Convolutional Network. *APPL SOFT COMPUT* 2019, 84:105685.

9. Liu T, Fan W, Wu C: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *ARTIF INTELL MED* 2019, 101:101723.

10. Giacalone M, Rasti P, Debs N, Frindel C, Cho TH, Grenier E, Rousseau D: Local spatio-temporal encoding of raw perfusion MRI for the prediction of final lesion in stroke. *MED IMAGE ANAL* 2018, 50:117-126.

11. Reboucas FP, Sarmento RM, Holanda GB, de Alencar LD: New approach to detect and classify stroke in skull CT images via analysis of brain tissue densities. *Comput Methods Programs Biomed* 2017, 148:27-43.

12. Sacco RL, Kasner SE, Broderick JP, Caplan LR, Connors JJ, Culebras A, Elkind MS, George MG, Hamdan AD, Higashida RT *et al*: An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *STROKE* 2013, 44(7):2064-2089.

13. Zerna C, Thomalla G, Campbell B, Rha JH, Hill MD: Current practice and future directions in the diagnosis and acute treatment of ischaemic stroke. *LANCET* 2018, 392(10154):1247-1256.

14. Pandian JD, Gall SL, Kate MP, Silva GS, Akinyemi RO, Ovbiagele BI, Lavados PM, Gandhi D, Thrift AG: Prevention of stroke: a global perspective. *LANCET* 2018, 392(10154):1269-1278.

15. Stevens RJ, Kothari V, Adler AI, Stratton IM: The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci (Lond)* 2001, 101(6):671-679.

16. Assmann G, Cullen P, Schulte H: Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *CIRCULATION* 2002, 105(3):310-315.

17. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U *et al*: Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *EUR HEART J* 2003, 24(11):987-1003.

18. Albarqouni L, Doust JA, Magliano D, Barr EL, Shaw JE, Glasziou PP: External validation and comparison of four cardiovascular risk prediction models with data from the Australian Diabetes, Obesity and Lifestyle study. *Med J Aust* 2019, 210(4):161-167.

19. van Staa TP, Gulliford M, Ng ES, Goldacre B, Smeeth L: Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLOS ONE* 2014, 9(10):e106455.

20. Xing X, Yang X, Liu F, Li J, Chen J, Liu X, Cao J, Shen C, Yu L, Lu F *et al*: Predicting 10-Year and Lifetime Stroke Risk in Chinese Population. *STROKE* 2019, 50(9):2371-2378.

21. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB: Probability of stroke: a risk profile from the Framingham Study. *STROKE* 1991, 22(3):312-318.

22. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB: Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *STROKE* 1994, 25(1):40-43.

23. Wan E, Fong D, Fung C, Yu E, Chin WY, Chan A, Lam C: Development of a cardiovascular diseases risk prediction model and tools for Chinese patients with type 2 diabetes mellitus: A population-based retrospective cohort study. *DIABETES OBES METAB* 2018, 20(2):309-318.

24. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, Negahban SN, Krumholz HM: Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circ Cardiovasc Qual Outcomes* 2016, 9(6):629-640.

25. Ngiam KY, Khor IW: Big data and machine learning algorithms for health-care delivery. *LANCET ONCOL* 2019, 20(5):e262-e273.

26. Doupe P, Faghmous J, Basu S: Machine Learning for Health Services Researchers. *VALUE HEALTH* 2019, 22(7):808-815.

27. Wiemken TL, Kelley RR: Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health* 2020, 41:21-36.

28. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, Wang Y, Douiri A, Wolfe CD, Bray B: A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLOS ONE* 2020, 15(6):e234722.

29. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G: Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *INT J EPIDEMIOL* 2014, 43(1):61-68.

30. He H, Garcia E: Learning from Imbalanced Data. *IEEE T KNOWL DATA EN* 2009.

31. Chawla N, Bowyer K, Hall L, Kegelmeyer W: SMOTE: Synthetic Minority Over-sampling Technique. *J ARTIF INTELL RES* 2002.

32. Stoppiglia H, Dreyfus G, Dubois R, Oussar Y: Ranking a Random Feature for Variable and Feature Selection. *J MACH LEARN RES* 2003.

33. Kursa MB, Rudnicki WR: Feature Selection with the Boruta Package. *J STAT SOFTW* 2010, 36(11).

34. Pencina MJ, D'Agostino RS, D'Agostino RJ, Vasan RS: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *STAT MED* 2008, 27(2):157-172, 207-212.

35. Kerr KF, McClelland RL, Brown ER, Lumley T: Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *AM J EPIDEMIOL* 2011, 174(3):364-374.

36. Breiman L: Random forests. *MACH LEARN* 2001.

37. Chen TQ, Guestrin C: XGBoost: A Scalable Tree Boosting System. 2016.

38. Vickers AJ: Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *AM STAT* 2008, 62(4):314-320.

39. Huang Y, Sullivan PM, Feng Z: Evaluating the predictiveness of a continuous marker. *BIOMETRICS* 2007, 63(4):1181-1188.

40. Lundberg S, Erion G, Lee S: Consistent feature attribution for tree ensembles. 2018.

41. Alexandropoulos SN, Kotsiantis SB, Vrahatis MN: Data preprocessing in predictive data mining. *The Knowledge Engineering Review* 2019, 34.

42. Benhar H, Idri A, Fernandez-Aleman JL: Data preprocessing for heart disease classification: A systematic literature review. *Comput Methods Programs Biomed* 2020, 195:105635.

43. Guyon I, Elisseeff A: An Introduction to Variable and Feature Selection. *The Journal of Machine Learning Research* 2003.

44. Rice ME, Harris GT: Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum Behav* 2005, 29(5):615-620.

45. Singal A, Mukherjee A, Elmunzer B, Higgins P, Waljee A: Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma. *AM J GASTROENTEROL* 2013.

46. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T, Chihara H, Fukumitsu R, Okawa M, Yamana N *et al*: Predicting Clinical Outcomes of Large Vessel Occlusion Before Mechanical Thrombectomy Using Machine Learning. *STROKE* 2019, 50(9):2379-2388.

47. Hatton CM, Paton LW, McMillan D, Cussens J, Gilbody S, Tiffin PA: Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *J Affect Disord* 2019, 246:857-860.

48. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK: Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA CARDIOL* 2017, 2(2):204-209.

49. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y: Integrating the predictiveness of a marker with its performance as a classifier. *AM J EPIDEMIOL* 2008, 167(3):362-368.

50. Boehme AK, Esenwa C, Elkind MS: Stroke Risk Factors, Genetics, and Prevention. *CIRC RES* 2017, 120(3):472-495.

51. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015, 350:g7594.

Tables

Table 1 Baseline characteristics of the study population

Category	Variable	Total	Stroke	Non_stroke
Demographics	Age (year)	66.75(62.92,72.56)	69.67(64.08,75.83)	66.75(62.83,72.5)
	Gender（man）	2983(50.05%)	71 (54.20%)	2912(49.96%)
	WHtR	0.54(0.50,0.58)	0.54(0.51,0.57)	0.54(0.50,0.58)
Lifestyle	Smoking (yes)	2398(41.91%)	58(44.27%)	2440(41.86%)
	Drinking (yes)	1817(30.49%)	34(25.95%)	1783(30.59%)
Disease history	Hypertension (yes)	1832(30.74%)	62(47.33%)	1770(30.37%)
	Diabetes (yes)	388(6.51%)	17(12.98%)	371(6.36%)
	Dyslipidemia (yes)	564(9.49%)	25(19.08%)	539(9.25%)
	Heart_disease (yes)	887(14.88%)	29(22.14%)	858(14.72%)
Clinical data	hsCRP (mg/L)	1.01(0.79,1.28)	1.01(0.95,1.34)	1.01(0.79,1.28)
	WBC (10^9/L)	5.90(5.39,6.50)	5.90(5.90,6.80)	5.90(5.30,6.50)
	GLU (mg/dL)	102.06(98.64,105.84)	102.06(101.16,107.1)	102.06(98.64,105.75)
	HbA1c (%)	5.10(5.00,5.30)	5.10(5.00,5.10)	5.10(5.00,5.30)
	LDLC (mg/dL)	115.59(103.22,127.87)	115.59(103.22,125.64)	115.59(103.22,127.96)
	TG (mg/dL)	100.89(84.96,116.82)	100.89(97.35,130.1)	100.89(84.96,116.82)
	CysC (mg/L)	1.06(1.02,1.10)	1.06(1.05,1.08)	1.06(1.02,1.10)

Table 2 IDI and NRI for prediction models

Model	IDI	*Z* value	*p*-value	NRI	*Z* value	*p*-value
LR	0.001	0.166	0.868	0.011	0.686	0.493
RF	-0.001	-0.190	0.849	-0.009	-0.685	0.493
XGBoost	0.006	0.293	0.770	0.052	0.914	0.361

LR, logistic regression; RF, random forest; XGBoost, extreme gradient boosting. IDI, integrated discrimination improvement; NRI, net reclassification improvement.

Table 3 Discrimination and calibration of prediction models

Model		Testing set		Accuracy	Sensitivity	Specificity	AUC (95%CI)	BS (95%CI)
		Stroke	Non-stroke					
LR	Stroke	21	450	0.738	0.525	0.726	0.718(0.649,0.787)	0.022(0.015,0.028)
	Non-stroke	19	1299					
RF	Stroke	30	450	0.743	0.750	0.726	0.823(0.759,0.886)	0.019(0.014,0.025)
	Non-stroke	10	1299					
XGBoost	Stroke	28	454	0.740	0.700	0.724	0.808(0.730,0.886)	0.020(0.015,0.026)
	Non-stroke	12	1295					

Table 4 Variable importance of prediction models

Variable	Logistic regression		Random forest		XGBoost	
	OR	Rank	Importance	Rank	Importance	Rank
Age	1.049	8	0.071	9	0.085	9
Gender	1.353	3	0.01	12	0.009	12
WHtR	1.345	4	0.076	8	0.096	5
Smoking	0.913	7	0.012	11	0.01	11
Dyslipidemia	2.06	1	0.026	10	0.037	10
hsCRP	1.041	9	0.08	6	0.089	7
WBC	1.279	5	0.136	3	0.111	3
GLU	1.028	10	0.147	2	0.15	1
HbA1c	0.648	2	0.163	1	0.133	2
LDLC	0.998	12	0.097	5	0.104	4
TG	1.005	11	0.101	4	0.089	6
CysC	1.118	6	0.08	7	0.086	8

Figures

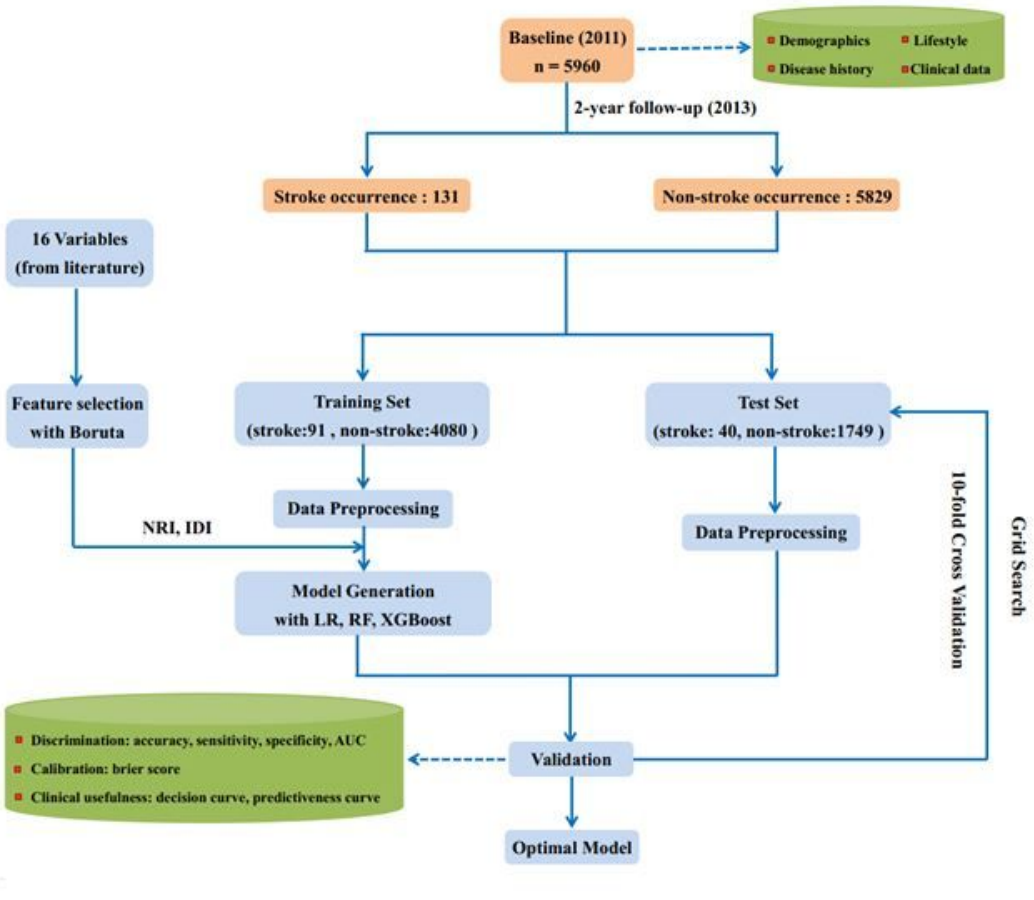

Figure 1

The flowchart of model derivation and internal validation. LR, logistic regression; RF, random forest; XGBoost, extreme gradient boosting; NRI, net reclassification improvement; IDI, integrated discrimination improvement; AUC, area under the receiver operating characteristic curve.

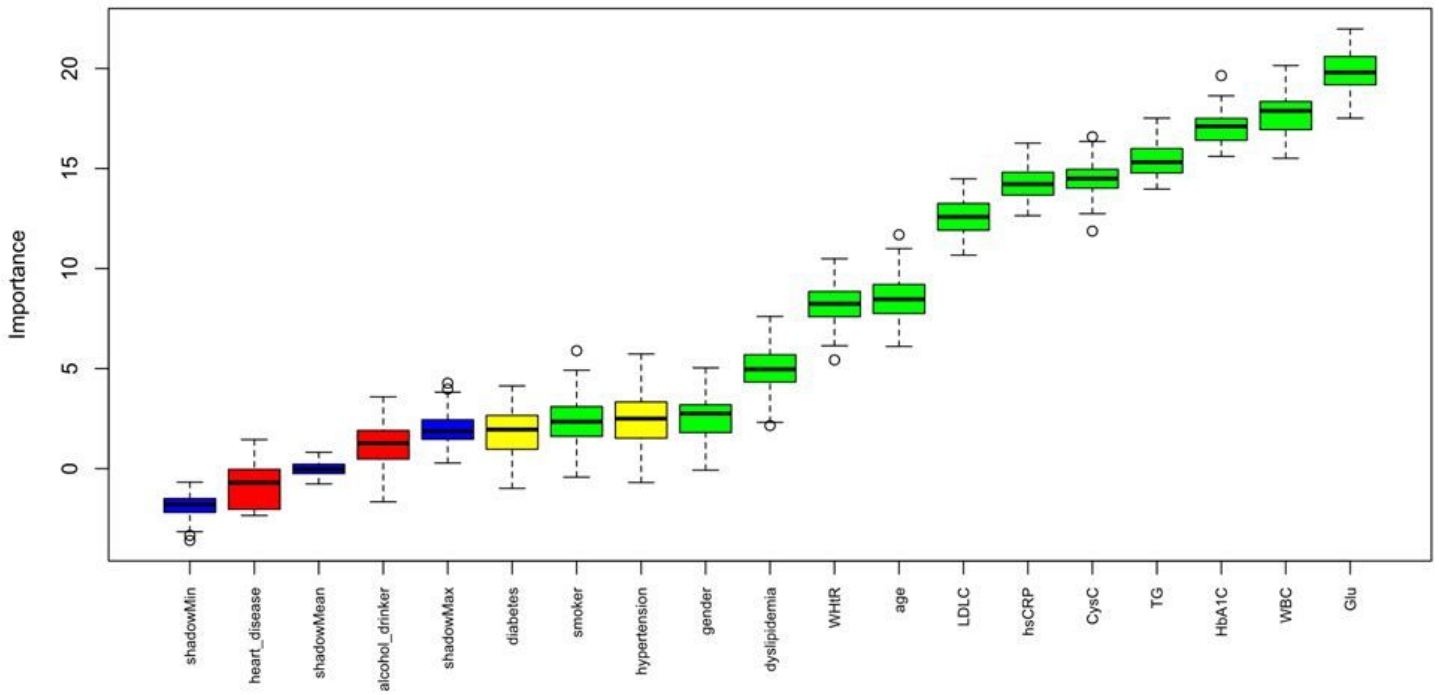

Figure 2

Results of feature selection with Boruta. WHtR, ratio between waist (cm) and height (cm). hsCRP, hypersensitive C-reactive protein; WBC, white blood cell; GLU, glucose; HbA1c, glycated hemoglobin; LDLC, LDL cholesterol; TG, triglycerides; CysC, cystatin C

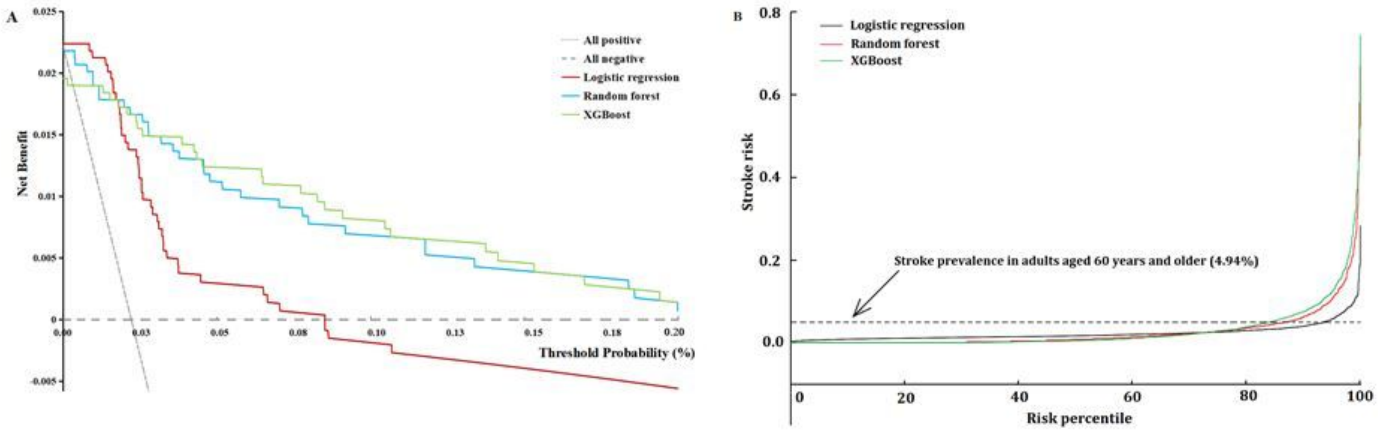

Figure 3

A. Decision curve for predicting stroke patients with traditional logistic regression and ensemble learning methods (random forest, and extreme gradient boosting). The x-axis indicates the threshold probability for outcome of stroke, the y-axis indicates the net benefit. Two extreme treatments, including all participants were regarded as stroke and all were regarded as non-stroke, were used as the references. B. Predictiveness curve for Logistic regression, random forest, and XGBoost (extreme gradient boosting).

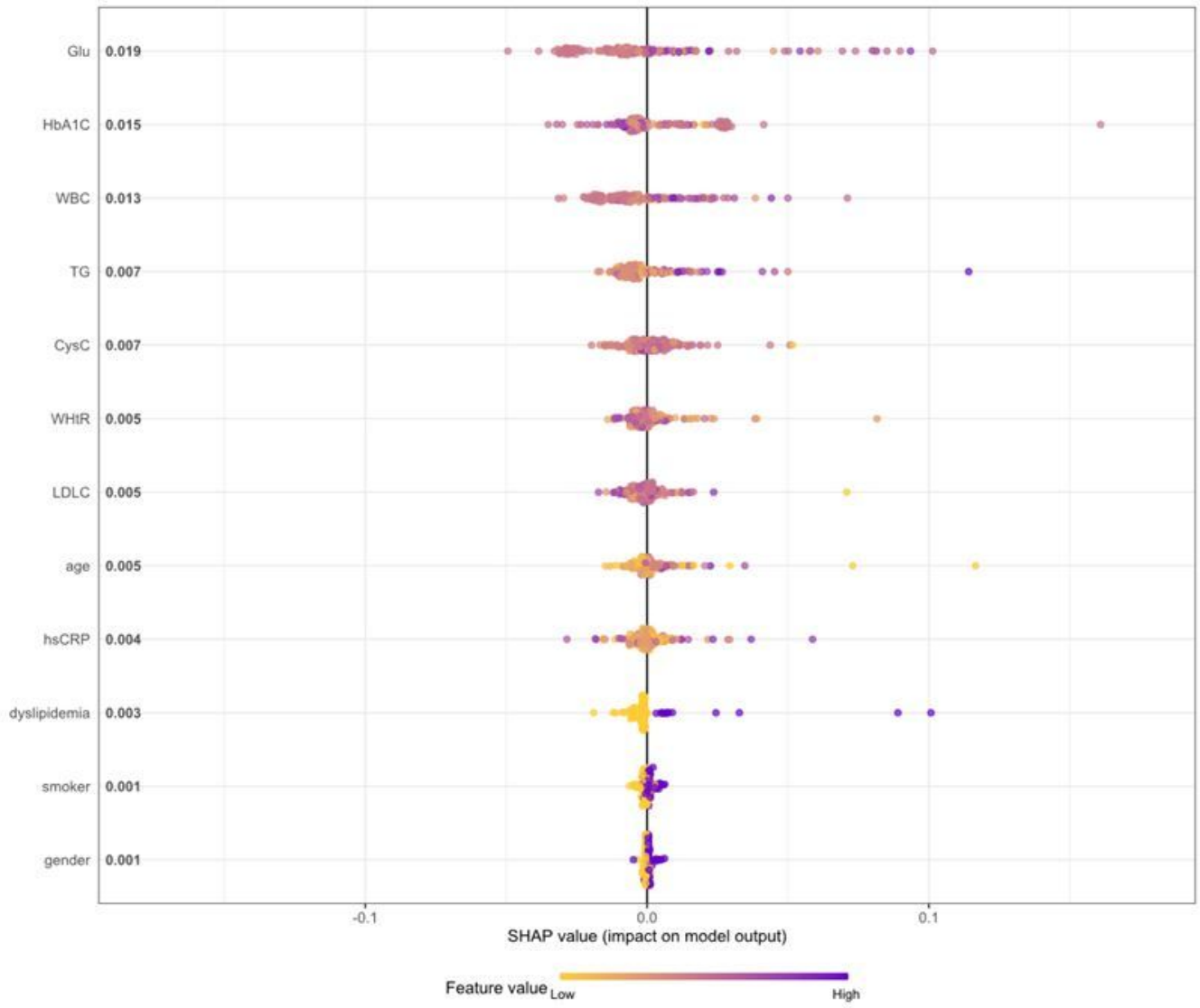

Figure 4

The SHAP value of each feature in XGBoost (extreme gradient boosting) and its importance. All the features were sorted in the descending order according to SHAP values.