BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Machine learning–based predictive model for post-stroke dementia

Zemin Wei[1], Mengqi Li[2], Chenghui Zhang[2], Jinli Miao[3], Wenmin Wang[3] and Hong Fan[1*]

Abstract

Background Post-stroke dementia (PSD), a common complication, diminishes rehabilitation efficacy and affects disease prognosis in stroke patients. Many factors may be related to PSD, including demographic, comorbidities, and examination characteristics. However, most existing methods are qualitative evaluations of independent factors, which ignore the interaction amongst various factors. Therefore, the purpose of this study is to explore the applicability of machine learning (ML) methods for predicting PSD.

Methods 9 acceptable features were screened out by the Spearman correlation analysis and Boruta algorithm. We developed and evaluated 8 ML models: logistic regression, elastic net, k-nearest neighbors, decision tree, extreme gradient boosting, support vector machine, random forest, and multilayer perceptron.

Results A total of 539 stroke patients were included in this study. Among the 8 models used to predict PSD, extreme gradient boosting and random forest showed the highest area under the curve (AUC) of the receiver operating characteristic curve (ROC), with values of 0.7287 and 0.7285, respectively. The most important features for predicting PSD included age, high sensitivity C-reactive protein, stroke side and location, and the occurrence of cerebral hemorrhage.

Conclusion Our findings suggest that ML models, especially extreme gradient boosting, can best predict the risk of PSD.

Keywords Stroke, Post-stroke dementia, Boruta algorithm, Machine learning, Prediction model

Introduction

Stroke, also known as cerebrovascular accident (CVA), stands as the second leading cause of death worldwide. It is characterized by a high incidence rate, elevated mortality, and increased disability rates [1]. According to the Global Burden of Disease study in 2019, there were 6.55 million fatalities attributed to stroke, with 12.2 million new cases reported. Notably, China ranks among the countries with the highest proportion of stroke risk factors globally [2]. Advances in medical treatments have notably reduced the mortality associated with stroke. It is noteworthy that post-stroke disability complications, post-stroke cognitive impairment (PSCI) persist widely. According to severity, PSCI is divided into post-stroke

*Correspondence:
Hong Fan
fanhongsx@126.com
[1]Department of Geriatrics, Shaoxing People's Hospital, Shaoxing, Zhejiang, P. R. China
[2]School of Medicine, Shaoxing University, Shaoxing, Zhejiang, P. R. China
[3]The Yangtze River Delta Biological Medicine Research and Development Center of Zhejiang Province, Yangtze Delta Region Institution of Tsinghua University, Hangzhou 314006, Zhejiang, P.R. China

cognitive impairment no dementia (PSCIND) and post-stroke dementia (PSD) [3].

According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), cognitive impairment can be classified into two categories based on severity: mild cognitive impairment (MCI) and dementia. Additionally, research has found that MCI is reversible and may progress back to a normal cognitive state [4]. About 20–30% of PSCI patients will deteriorate and develop into PSD [5, 6]. Thus, early diagnosis and treatment of PSCI and PSD contribute to improving the prognosis of stroke patients, alleviating societal and economic burdens. Numerous studies have reported many independent risk factors for PSCI and PSD related to age, hypertension, high cholesterol, diabetes, smoking, atrial fibrillation [7–9]. Stroke history and vascular risk factors (such as high blood pressure, obesity, and smoking) may accelerate the development of PSCI and PSD [10]. Thus, clinicians face challenges in integrating multiple factors for the early diagnosis of PSCI and PSD. The CHANGE and SIGNAL2 scales have demonstrated accuracy in predicting PSCI, with area under the curve (AUC) of receiver operating characteristic curve (ROC) ranging from 0.740 to 0.829 [11, 12]. Due to the complexity of the pathophysiology and multiple contributing factors associated with PSCI, conventional scoring systems with limited variables may not optimally predict PSCI. Researchers have applied cognitive impairment risk prediction models to PSCI, but the predictive performance has not been ideal [13].

Machine learning (ML) algorithms are adaptable to various types and sizes of data and have garnered significant attention in the development of patient-centered prediction/prognosis models. These models can help optimize treatment plans and facilitate the monitoring and management of health conditions. Recently, ML has demonstrated immense potential in enhancing the speed and accuracy of stroke imaging assessments [14, 15]. Consequently, some researchers are pivoting their focus towards ML, aiming to develop a more accurate predictive model for PSCI [16–18]. In this context, we conducted a retrospective study focusing on a more severe form of cognitive decline after stroke, PSD, with the goal of developing and testing the applicability of ML models in predicting PSD.

Materials and methods

Data source

For this retrospective cohort study, we selected 545 stroke inpatients from the rehabilitation department and geriatrics department of Shaoxing People's Hospital, spanning from January 2019 to August 2021.

Inclusion Criteria: (1) admitted for stroke between January 2019 and August 2021; (2) aged 18 years and above; (3) signed an informed consent form.

Exclusion Criteria: (1) Pre-existing cognitive impairment before the current stroke; (2) Presence of other diseases severely affecting cognitive function, such as anxiety, depression, or brain tumors.

Clinical variables

A total of 46 variables were collected, including demographic data, vascular risk factors, and examination findings. Demographic data included variables such as gender, age, occupation, and education level, and Body mass index (BMI). The vascular risk factors recorded were smoking status, alcohol consumption, hypertension, hyperlipidemia, and any history of stroke. Clinical examination findings compiled for analysis included the type of stroke, fibrinogen levels, high-sensitivity C-reactive protein (CRP), blood uric acid levels, blood homocysteine levels (Hcy), and specifics regarding the stroke's side and location.

Model construction and verification

The initial sample ($n=545$) underwent multicollinearity assessment to ensure the stability of subsequent analyses. Based on Spearman correlation analysis, we identified and removed individual significant and highly correlated features to mitigate the effects of multicollinearity. Additionally, features with over 20% missing values (BMI) were discarded, and 6 samples with extreme deviations and missing data were excluded. 539 patients and 45 variables were included in the study. Supplementary Table 1 shows the detailed information of all variables. The continuous variables in the remaining sample were standardized through Z-score normalization. Finally, the data of 539 patients were randomly divided into a training set and a test set in a 3:1 ratio.

In this study, we preselected features for the training set using the Boruta algorithm, a feature selection method based on the random forest algorithm [19]. The Boruta algorithm is a highly popular and robust feature selection method that helps in retaining only the most statistically significant variables [20], thereby making a practical contribution to our model. It identifies crucial features by creating shadow features (randomized copies of features) and evaluating the importance of original features using the random forest algorithm. Subsequently, we input these preselected features into 8 different ML models: logistic regression (LR), elastic net (EN), k-nearest neighbors (KNN), decision tree (DT), extreme gradient boosting (XGB), support vector machine (SVM), random forest (RF), and multilayer perceptron (MLP). For each model, we selected a set of hyperparameters that maximized the AUC of ROC on the training set using

a Bayesian optimizer, ensuring optimal performance and effective prediction and comparison on the test set. Detailed hyperparameters are available in the supplementary material. All models were subjected to five-fold cross-validation to ensure robustness and reliability. We plotted the ROC of each model on the test set to evaluate their predictive performance and decision curves analysis (DCA) and clinical impact curve (CIC) to assess their clinical utility.

Comparing models, we selected the best model based on AUC values and created a Shapley Additive exPlanations (SHAP) explainer to calculate SHAP values, which indicate the contribution of each feature to the prediction outcome. We plotted a SHAP summary chart to illustrate the impact of model features.

Statistical analysis

R software (version 4.3.0, https://www.R-project.org/) was used to support grouping and data statistical analysis. Continuous variables were presented as median with interquartile range (Mean, IQR), while categorical data were represented by numerical values and corresponding percentages (n, %). We used t-test or the Wilcoxon rank-sum non-parametric test for continuous variables, and chi-square test for categorical variables to compare the demographic and clinical characteristics between the PSD group and the non-PSD group.

Results

Characteristics of patients

Among the 539 stroke patients included in this analysis, 194 did not develop PSD, while 345 did. The baseline characteristics of the PSD and no PSD groups are shown in Table 1. There were significant differences in age and High-sensitive CRP between the two groups ($P<0.001$).

Model building and verification

Through Spearman correlation analysis, we removed 6 individual significant and highly correlated features. Feature selection was then conducted using the Boruta algorithm, identifying 9 acceptable variables: brain stem, age, temporal lobe, right lesion, cerebral hemorrhage, high-sensitive CRP, subarachnoid space, outer capsule, and Island leaves (Fig. 1). Using these 9 variables, we developed 8 ML models to predict the risk of PSD in stroke patients. Fig. 2 displays the ROC of all models, with their predictive discrimination represented by the AUC. The XGB model had the highest AUC (0.7287), followed by RF (0.7285), KNN (0.7113), MLP (0.7082), EN (0.7033), LR (0.7022), DT (0.6502), and SVM (0.6098). The clinical applicability of all models was further assessed using DCA and CIC. DCA was utilized to evaluate the clinical benefit of the predictive models. The threshold range for the XGB model was approximately 0-0.87, slightly narrower than those for the EN, LR, and KNN models. Additionally, the DCA overlapped partially or entirely with other models across most threshold ranges, indicating no significant difference in net benefits. However, in the threshold range of approximately 0.76–0.8, the net benefit of the XGB model was significantly higher than that of other models (Fig. 3). Supplementary Fig. 1 shows the CIC for all models, assessing the efficiency of the models. When the threshold was greater than 0.7, the high-risk PSD group identified by the XGB prediction model closely matched the actual PSD occurrences, confirming the model's high clinical efficiency. Additionally, we calculated the accuracy, sensitivity, and specificity for all models (Table 2). The XGB model achieved the highest accuracy with a value of 0.72939. Although the sensitivity and specificity of the XGB model were not the highest, its overall performance is better when these metrics are considered comprehensively.

Feature importance

To visually present the selected variables, we analyzed the best predictive model, XGB, using the SHAP package to show the positive and negative impact of each feature on a given sample. Fig. 4A displays the absolute values of the average SHAP values for different features. Age had the most significant impact on model output, followed by high-sensitive CRP, right lesion, temporal lobe, cerebral hemorrhage, brain stem, undergraduate, bilateral lesions, and subarachnoid space. Fig. 4B provides a more detailed view of the impact of each feature on individual predictions.

Discussion

In this retrospective cohort study, we selected 9 acceptable features using the Boruta algorithm and developed 8 ML models to predict the risk of PSD in stroke patients. Among the 8 models, the XGB model showed the highest AUC and good clinical applicability. Furthermore, the most impactful features for predicting PSD, in descending order of importance, were age, high-sensitive CRP, right lesion, temporal lobe, and cerebral hemorrhage.

Increasingly, research acknowledges that the risk factors and progression of post-stroke cognitive impairment and dementia (PSCID) are determined by a multitude of factors, including age, comorbidities, type of stroke (ischemic and hemorrhagic), education level, and the location and size of the stroke [3, 21]. Yan et al.'s investigation, which developed 8 model for predicting the occurrence of MCI after stroke, revealed that the LR model achieved the highest AUC of 0.8595. It also exhibited high accuracy, sensitivity, and specificity, at 0.770, 0.778, and 0.765 respectively [18]. However, owing to the small sample size ($n=199$), the model may not be sufficiently effective. Therefore, more patients were included

Table 1 Clinical characteristics of patients ($n = 539$)

Variables	Dementia		
	NO ($n = 194$)	Yes ($n = 345$)	*P*-vaule
Age (year)	66 (56, 72)	70 (60, 77)	< 0.001
Gender			0.052
Female	60 (30.93%)	137 (39.71%)	
Male	134 (69.07%)	208 (60.29%)	
Occupation			
Farmer	66 (34.02%)	133 (38.55%)	0.341
Worker	8 (4.12%)	11 (3.19%)	0.748
Staff	53 (27.32%)	75 (21.74%)	0.175
Merchant	11 (5.67%)	24 (6.96%)	0.689
Retire	56 (28.87%)	102 (29.57%)	0.942
Education			
Elementary school or below	74 (38.14%)	127 (36.81%)	0.83
Junior high school	57 (29.38%)	87 (25.22%)	0.344
High school	14 (7.22%)	35 (10.14%)	0.328
Undergraduate	17 (8.76%)	17 (4.93%)	0.116
Master degree or above	0 (0%)	1 (0.29%)	1
Smoking	54 (27.84%)	86 (24.93%)	0.524
Drinking	54 (27.84%)	91 (26.38%)	0.791
Hypertension	149 (76.8%)	272 (78.84%)	0.66
Hyperlipidemia	68 (35.05%)	127 (36.81%)	0.753
Diabetes	72 (37.11%)	136 (39.42%)	0.663
Previous Stroke	24 (12.37%)	49 (14.2%)	0.642
Accompanied by Depression	25 (12.89%)	46 (13.33%)	0.988
Cerebral hemorrhage	130 (67.01%)	210 (60.87%)	0.185
Cerebral infarction	54 (27.84%)	135 (39.13%)	0.185
Lesion			
Left	98 (52.06%)	121 (35.07%)	< 0.001
Right	71 (36.6%)	167 (48.41%)	0.01
Bilateral	25 (12.89%)	57 (16.52%)	0.316
Stroke Site			
Basal ganglia	100 (51.55%)	194 (56.23%)	0.338
Frontal lobe	26 (13.4%)	98 (28.41%)	< 0.001
Parietal lobe	29 (14.95%)	81 (23.48%)	0.025
Thalamus	12 (6.19%)	27 (7.83%)	0.594
Lateral ventricle	58 (29.9%)	68 (19.71%)	0.01
Subarachnoid space	0 (0%)	8 (2.32%)	0.077
Occipital lobe	8 (4.12%)	31 (8.99%)	0.055
Brain stem	25 (12.89%)	13 (3.77%)	< 0.001
Semi-oval area	17 (8.76%)	31 (8.99%)	1
Outer capsule	0 (0%)	6 (1.74%)	0.156
Inner capsule	2 (1.03%)	9 (2.61%)	0.354
Cerebellum	13 (6.7%)	15 (4.35%)	0.327
Island leaves	1 (0.52%)	13 (3.77%)	0.046
Radiation crown	2 (1.03%)	2 (0.58%)	0.95
Corpus callosum	3 (1.55%)	6 (1.74%)	1
Temporal lobe	28 (14.43%)	97 (28.12%)	< 0.001
Fibrinogen (g/L)	3.35 (2.88, 3.95)	3.35 (2.88, 4.16)	0.452
High-sensitive CRP (mg/L)	3 (0.95, 8.94)	4.99 (1.63, 17.46)	< 0.001
Blood uric acid (μmol/L)	291.35 (222.58, 345.48)	284.4 (221.1, 346.1)	0.634
Blood homocysteine (μmol/L)	11.17 (9.41, 13.82)	11.17 (9.01, 13.27)	0.220

Notes: Continuous variables were presented as median with interquartile range (Mean, IQR), while categorical data were represented by numerical values and corresponding percentages (n, %)

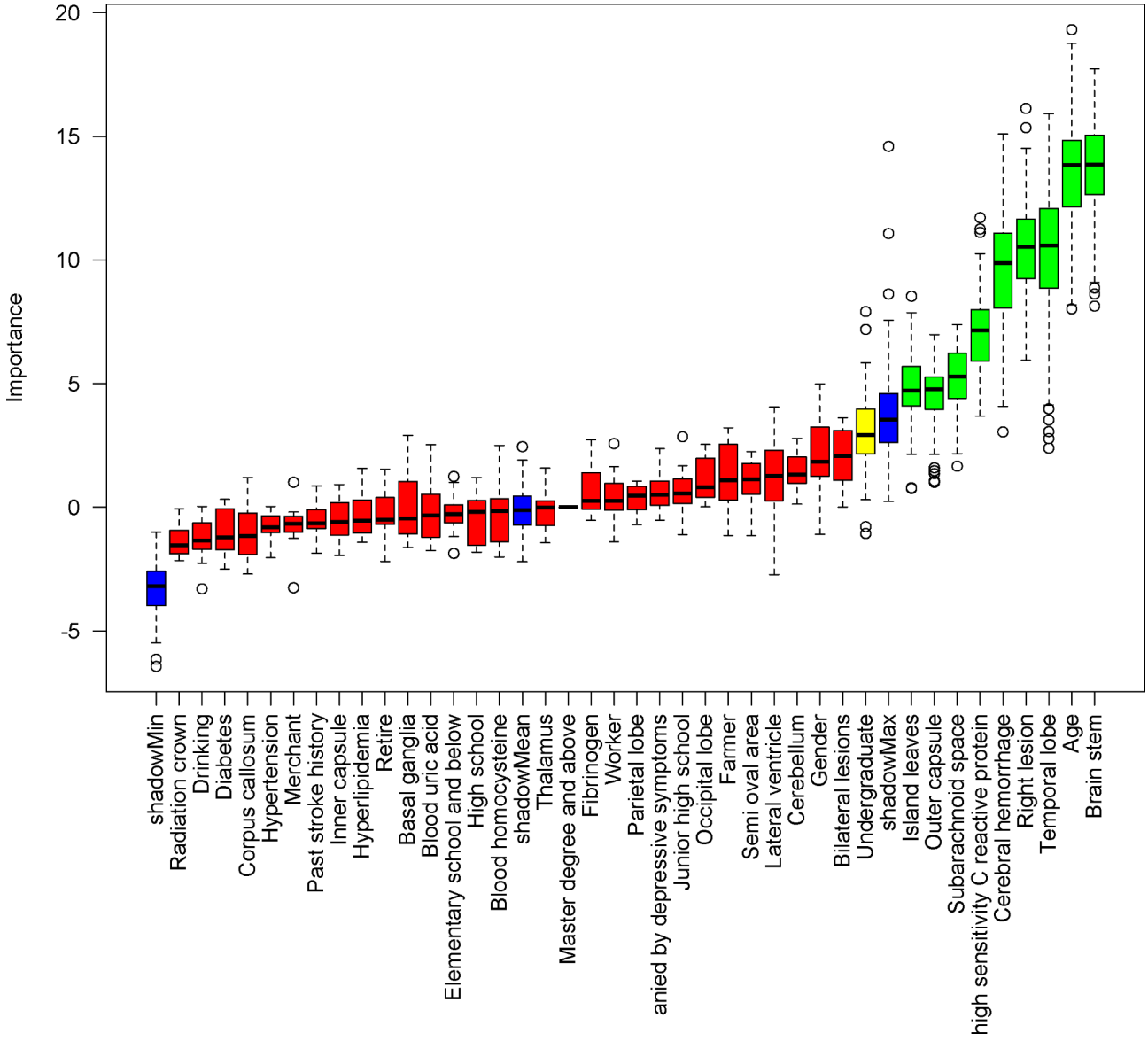

Fig. 1 Feature selection based on Boruta algorithm. Green represents acceptable variables

in this study to provide better confidence and stability. Given the complex and varied factors influencing PSCID, prediction in clinical settings is challenging, with lower accuracy compared to post-stroke functional outcomes. ML methods are often more suitable when dealing with complex influencing factors. Stroke can be categorized into hemorrhagic strokes (HS) and ischemic strokes (IS), with IS constituting a significantly higher proportion of all strokes, approximately 87% according to data from Johns Hopkins Medical Center [22]. Thus, research focus and resources are more inclined towards IS. Our study encompassed patients with both types of strokes, offering a more comprehensive perspective for understanding and predicting PSD. The studies of Lee et al. (n=951) and Ji et al. (n=397) developed a variety of ML models to predict the risk of PSCI in patients with acute IS (AIS). They found that the XGB and Gaussian Naive Bayes (GNB) models had the best discrimination, with AUC of 0.7919 and 0.925, respectively. Additionally, these models also surpassed the LR model in other metrics, including accuracy and F1 score [16, 17]. In our study, XGB also showed better overall performance for PSD prediction. Although PSCI and PSD may clinically overlap, they differ in definition, severity, treatment strategies, and management. Dementia represents a more severe clinical outcome, significantly impacting patients' daily lives and independence. Therefore, by concentrating on PSD as an outcome, effectively predicting its occurrence can facilitate the implementation of prevention and intervention measures at an earlier stage.

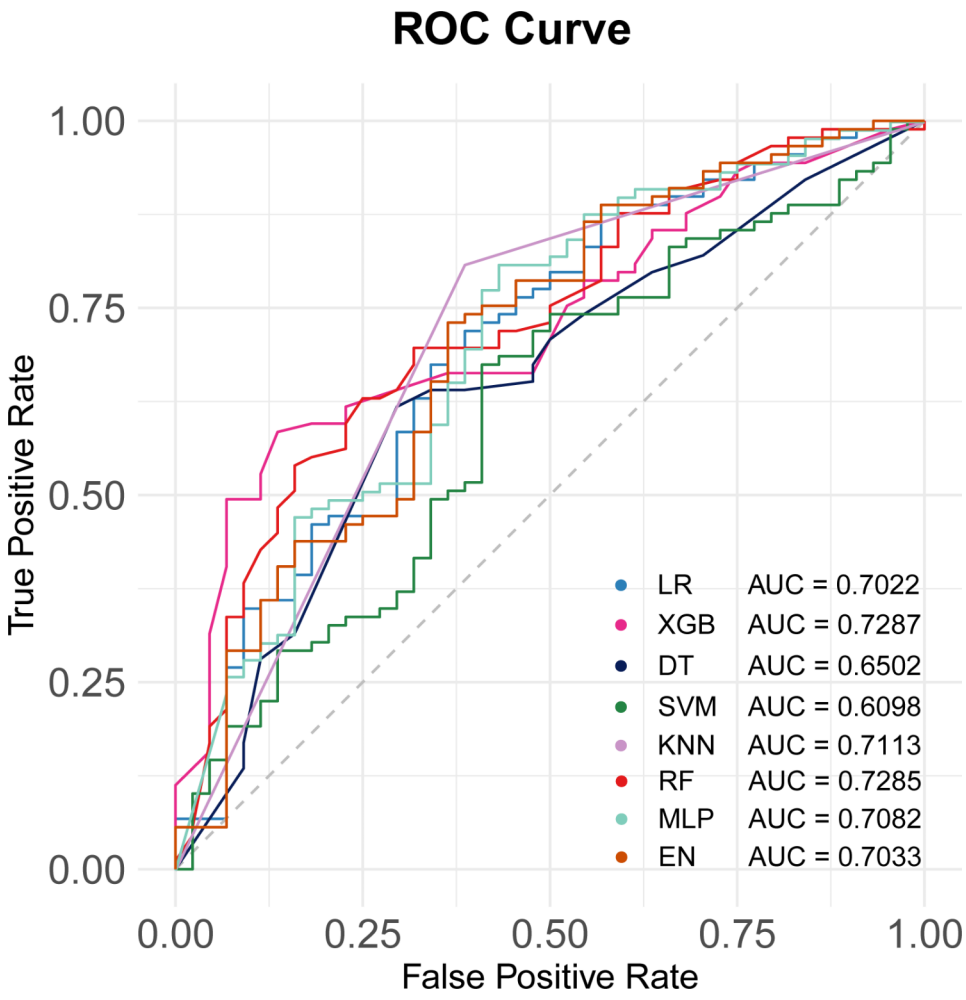

Fig. 2 The receiver operating characteristic curves of the eight machine learning models. LR, logistic regression; XGB, extreme gradient boosting; DT, decision tree; SVM, support vector machine; KNN, k-nearest neighbors; RF, random forest; MLP, multilayer perceptron; EN, elastic net

Despite the SIGNAL2 risk score and CHANGE risk score models exhibiting good discriminative ability in predicting PSCI, both models use Mini-Mental Status Examination (MMSE) ≤ 25 or Montreal Cognitive Assessment (MOCA) ≤ 22 as cut-off raw scores and incorporate age and education level as variables with high weight. However, since MMSE and MOCA scores are highly dependent on age and education level [11, 12]. Therefore, regardless of the clinical characteristics of the patient, it may have little effect on the predictive outcome. In our study, the ML models we constructed can continually learn and adapt to new data, improving their accuracy over time, unlike the static SIGNAL2 and CHANGE models. By applying the Boruta algorithm, we identified the most important features in the dataset for the predictive models. This approach, unconstrained by data type, allows for comprehensive feature selection, eliminating irrelevant features to reduce the risk of overfitting, thereby enhancing the model's accuracy and interpretability. We incorporated variables in the green zone into the 8 models, with the XGB model achieving the highest AUC value. Lastly, we employed SHAP to quantify the contribution of each feature to the predictive outcome. In the XGB model, the most important features were age, high-sensitive CRP, right lesion, temporal lobe, and cerebral hemorrhage. These key features show both consistency and differences with previous research on PSCID risk factors [16–18, 23].

Most studies agree on the close association between age and cognitive decline. For instance, the REGARDS study found that each additional year of baseline age increased the likelihood of cognitive impairment by 17% during the follow-up period [23], aligning with non-stroke population studies that identified older age as a significant risk factor for cognitive impairment [24]. However, Yan et al.'s findings suggested no correlation between age and the occurrence of MCI after stroke [18]. A systematic review and meta-analysis assessed the potential of various blood-derived proteins as biomarkers for PSCI, recommending Hcy, CRP, total cholesterol, and low-density

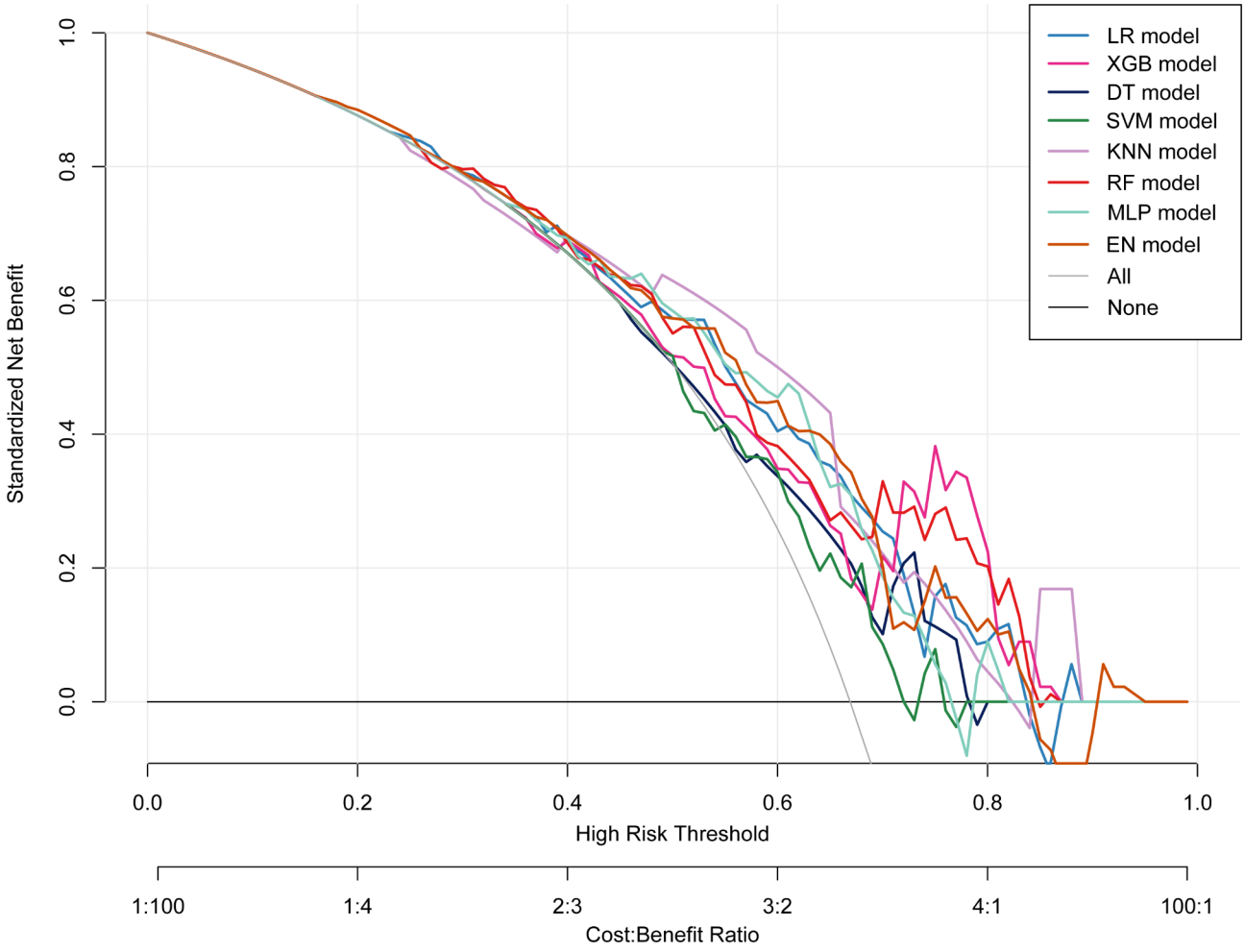

Fig. 3 The DCA curves of the eight machine learning models. LR, logistic regression; XGB, extreme gradient boosting; DT, decision tree; SVM, support vector machine; KNN, k-nearest neighbors; RF, random forest; MLP, multilayer perceptron; EN, elastic net

Table 2 Performances of various prediction models predicting PSD using a testing data set

	Accuracy	Sensitivity	Specificity
DT	0.6391	0.5000	0.7079
ENET	0.6992	0.1136	0.9888
XGB	0.7293	0.6136	0.7865
Logistic	0.6917	0.4773	0.7978
MLP	0.7143	0.5227	0.8090
RF	0.7068	0.4091	0.8539
SVM	0.3383	0.6591	0.1798
KNN	0.6767	0.3182	0.8539

lipoprotein as potential biomarkers for PSCI [25]. The high-sensitivity CRP test, capable of measuring low concentrations of CRP in blood, is useful for assessing low-grade inflammation and cardiovascular risk. A significant association between high-sensitive CRP concentrations and long-term cognitive decline was observed in a large study involving 5257 participants [26], consistent with our findings. Our research provides additional evidence supporting high-sensitive CRP as a potential biomarker for PSD, enhancing its potential in PSD prediction and monitoring. No association between Hcy and PSD was found in our study, although high Hcy levels have been confirmed as a risk factor for cerebrovascular events and cognitive decline [25]. The reasons for these differences may include sample selection bias, differences in data collection, or analytical methods.

At present, ML prediction models related to PSCID primarily focus on PSCI. To the best of our knowledge, this study is the first to construct and compare the performance of eight different risk prediction models for PSD. Furthermore, this research integrates ML techniques with demographic and imaging features to predict PSD. However, there are several limitations to our study that cannot be ignored. Firstly, the dataset used in this study was sourced from patients in the geriatrics department and those transferred to the rehabilitation department, who may exhibit significant differences in distributions of various demographic characteristics,

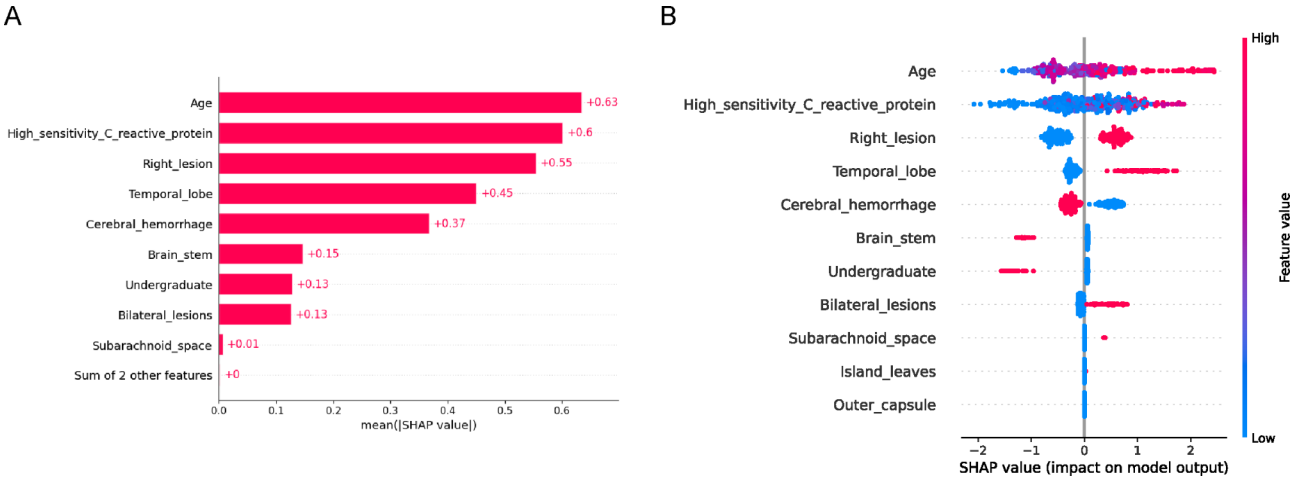

Fig. 4 The Shapley Additive exPlanations values of the best prediction model, XGB. (**A**) Average impact of features on model predictions. (**B**) Detailed impact analysis of each feature

such as age and gender, compared to the general stroke patient population. These disparities could potentially limit the generalizability of the model across different demographic groups. Secondly, this is a single-center retrospective cohort study, and the data quality and diversity might be affected, necessitating external validation and optimization. Thirdly, not all patients underwent all examinations, leading to missing features in some cases. Although we excluded patients with more than 20% missing values and employed multiple imputation for features with less than 10% missing values to mitigate this concern, the possibility of residual effects remains. Finally, our model has overlooked social, psychological, and behavioral elements like social support and lifestyle, which are crucial for understanding cognitive outcomes after a stroke. Including these factors in future research could enhance the model's ability.

Conclusion

We proved that the ML model, especially the XGB model, can accurately predict PSD and is expected to be an effective assistant tool for the diagnosis and treatment of PSD. Among the variables included, age and high-sensitive CRP are the two most significant factors influencing the XGB model's output. However, the efficacy of this model in external cohorts and its potential to mitigate the occurrence of PSD remains to be determined.

Abbreviations

PSD	Post-stroke dementia
PSCI	Post-stroke cognitive impairment
ML	Machine learning
IS	Ischemic strokes
ROC	Receiver operating characteristic curve
AUC	Area under the curve
CRP	C-reactive protein
LR	Logistic regression
EN	Elastic net
KNN	k-nearest neighbors
DT	Decision tree
XGB	Extreme gradient boosting
SVM	Support vector machine
RF	Random forest
MLP	Multilayer perceptron
DCA	Decision curves analysis
CIC	Clinical impact curve
SHAP	Shapley Additive exPlanations
Hcy	Homocysteine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02752-4.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

Thanks to Engineer Hao Sun of the Yangtze River Delta Biological Medicine Research and Development Center of Zhejiang Province for checking and reviewing the 8 model codes: Logistic regression (LR), elastic Network (EN), K-nearest neighbor (KNN), Decision tree (DT), Extreme gradient boosting (XGB), Support vector Machine (SVM), Random Forest (RF) and Multilayer Perceptron (MLP).

Author contributions

Hong Fan: Funding acquisition, Conceptualization, Methodology, Data Curation, Roles/Writing-original draft. Zemin Wei and MengQi Li : Conceptualization, Methodology, Validation, Investigation, Formal analysis. Chenghui Zhang and Jinli Miao : Methodology, Data curation, Software, Formal analysis. Wenmin Wang and Muhammad Arshad : Conceptualization, Methodology, Investigation, Project administration, Writing-Review & Editing.

Funding

This work was funded by the Public Welfare Technology Application Research Project of Zhejiang Province (#LGF20H170005), the Traditional Chinese Medicine Scientific Research Fund Project of Zhejiang Province (#2020ZB270), the Health Science and Technology Program of Zhejiang Province (#2022KY1285), and Shaoxing Health Science and Technology Plan Project (# 2023SKY029). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
The datasets generated and/or analysed during the current study are not publicly available as there was no data sharing as part of the ethics approval (and raw data is potentially re-identifiable) but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This research has been performed in accordance with the Declaration of Helsinki and approved by Ethics Committee of Shaoxing People's Hospital (NO.2019-KK-Y-120-01). Written informed consent was obtained from individual or guardian participants. All methods were carried out in accordance with relevant guidelines and regulations in the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 March 2024 / Accepted: 4 November 2024
Published online: 11 November 2024

References

1. Katan M, Luft A. Global Burden of Stroke. Semin Neurol. 2018;38(2):208–11. https://doi.org/10.1055/s-0038-1649503.
2. Gorelick PB. The global burden of stroke: persistent and disabling. Lancet Neurol. 2019;18(5):417–8. https://doi.org/10.1016/s1474-4422(19)30030-4.
3. Rost NS, Brodtmann A, Pase MP, et al. Post-stroke Cognitive Impairment and Dementia. Circ Res. 2022;130(8):1252–71. https://doi.org/10.1161/circresaha.122.319951.
4. Shimada H, Doi T, Lee S, et al. Reversible predictors of reversion from mild cognitive impairment to normal cognition: a 4-year longitudinal study. Alzheimers Res Ther. 2019;11(1):24. https://doi.org/10.1186/s13195-019-0480-5.
5. Pendlebury ST, Rothwell PM. Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. Lancet Neurol. 2009;8(11):1006–18. https://doi.org/10.1016/s1474-4422(09)70236-4.
6. Mijajlović MD, Pavlović A, Brainin M, et al. Post-stroke dementia - a comprehensive review. BMC Med. 2017;15(1):11. https://doi.org/10.1186/s12916-017-0779-7.
7. Kaddumukasa MN, Kaddumukasa M, Katabira E, et al. Prevalence and predictors of post-stroke cognitive impairment among Stroke survivors in Uganda. Res Sq. 2023. https://doi.org/10.21203/rs.3.rs-2456615/v1.
8. Lo JW, Crawford JD, Desmond DW, et al. Profile of and risk factors for post-stroke cognitive impairment in diverse ethnoregional groups. Neurology. 2019;93(24):e2257–71. https://doi.org/10.1212/wnl.0000000000008612.
9. Sadhukhan D, Biswas A, Mishra S, et al. Genetic variations and altered blood mRNA level of circadian genes and BDNF as risk factors of Post-stroke Cognitive Impairment among Eastern indians. Neuromolecular Med. 2023. https://doi.org/10.1007/s12017-023-08761-2.
10. Sahathevan R, Brodtmann A, Donnan GA. Dementia, stroke, and vascular risk factors; a review. Int J Stroke. 2012;7(1):61–73. https://doi.org/10.1111/j.1747-4949.2011.00731.x.
11. Chander RJ, Lam BYK, Lin X, et al. Development and validation of a risk score (CHANGE) for cognitive impairment after ischemic stroke. Sci Rep. 2017;7(1):12441. https://doi.org/10.1038/s41598-017-12755-z.
12. Kandiah N, Chander RJ, Lin X, et al. Cognitive impairment after mild stroke: development and validation of the SIGNAL2 risk score. J Alzheimers Dis. 2016;49(4):1169–77. https://doi.org/10.3233/jad-150736.
13. Tang EYH, Price CI, Robinson L, et al. Assessing the predictive validity of simple dementia risk models in harmonized stroke cohorts. Stroke. 2020;51(7):2095–102. https://doi.org/10.1161/strokeaha.120.027473.
14. Mouridsen K, Thurner P, Zaharchuk G. Artificial Intelligence Applications in Stroke. Stroke. 2020;51(8):2573–9. https://doi.org/10.1161/strokeaha.119.027479.
15. Dritsas E, Trigka M. Stroke risk prediction with machine learning techniques. Sens (Basel). 2022;22(13). https://doi.org/10.3390/s22134670.
16. Ji W, Wang C, Chen H, et al. Predicting post-stroke cognitive impairment using machine learning: a prospective cohort study. J Stroke Cerebrovasc Dis. 2023;32(11):107354. https://doi.org/10.1016/j.jstrokecerebrovasdis.2023.107354.
17. Lee M, Yeo NY, Ahn HJ, et al. Prediction of post-stroke cognitive impairment after acute ischemic stroke using machine learning. Alzheimers Res Ther. 2023;15(1):147. https://doi.org/10.1186/s13195-023-01289-4.
18. Yan FJ, Chen XH, Quan XQ, et al. Development and validation of an interpretable machine learning model-Predicting mild cognitive impairment in a high-risk stroke population. Front Aging Neurosci. 2023;15:1180351. https://doi.org/10.3389/fnagi.2023.1180351.
19. Kursa MB, Rudnicki WRJJoSS. Feature Selection Boruta Package. 2010;36(11):1–13.
20. Zhou H, Xin Y, Li S. A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics. 2023;24(1):224. https://doi.org/10.1186/s12859-023-05300-5.
21. Filler J, Georgakis MK, Dichgans M. Risk factors for cognitive impairment and dementia after stroke: a systematic review and meta-analysis. Lancet Healthy Longev. 2024;5(1):e31–44. https://doi.org/10.1016/s2666-7568(23)00217-9.
22. Johns Hopkins Medicine. Types of Stroke. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/types-of-stroke. Accessed 4 Jan, 2023.
23. Levine DA, Wadley VG, Langa KM, et al. Risk factors for Poststroke Cognitive decline: the REGARDS Study (reasons for Geographic and racial differences in stroke). Stroke. 2018;49(4):987–94. https://doi.org/10.1161/strokeaha.117.018529.
24. Samson AD, Shen K, Grady CL, et al. Exploration of salient risk factors involved in mild cognitive impairment. Eur J Neurosci. 2022;56(9):5368–83. https://doi.org/10.1111/ejn.15665.
25. Kim KY, Shin KY, Chang KA. Potential biomarkers for Post-stroke Cognitive Impairment: a systematic review and Meta-analysis. Int J Mol Sci. 2022;23(2). https://doi.org/10.3390/ijms23020602.
26. Zheng F, Xie W. High-sensitivity C-reactive protein and cognitive decline: the English Longitudinal Study of Ageing. Psychol Med. 2018;48(8):1381–9. https://doi.org/10.1017/s0033291717003130.

Publisher's note