Available Online: 1 June 2026 DOI: 10.54254/3049-5458/2026.34050

Machine learning explainability-based investigation of the influencing factors of cerebral small vessel disease in patients with carotid plaque

Boyan Jia[1], Qiufen Shu[1], Cuicui Wu[1], Yeting Wang[1], Lizhen Li[2*]

[1]Graduate School, Youjiang Medical University for Nationalities, Baise, China
[2]Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, China

*Corresponding Author. Email: 85621508@qq.com

Abstract. Objective: To investigate the key influencing factors and complex interaction patterns associated with Cerebral Small Vessel Disease (CSVD) in patients with carotid plaque, and to construct a high-accuracy and highly interpretable CSVD risk prediction model. Methods: A retrospective case–control study was conducted, enrolling 373 patients with carotid plaque (207 CSVD-positive and 166 CSVD-negative cases). Demographic data, inflammatory biomarkers, and lipid profile indicators were collected. Least Absolute Shrinkage and Selection Operator (LASSO) regression was used for feature selection, and multivariate logistic regression was applied to identify independent risk factors. Restricted Cubic Spline (RCS) analysis was performed to assess potential nonlinear relationships between continuous variables and CSVD risk. Four machine learning algorithms—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Partial Least Squares Discriminant Analysis (PLS-DA), and XGBoost—were used to develop predictive models. The optimal model was further interpreted using the SHapley Additive exPlanations (SHAP) framework to provide both global and local explanations. Results: Multivariate logistic regression showed that age (OR = 1.09), Neutrophil count (N) (OR = 1.87), Low-Density Lipoprotein (LDL) (OR = 1.78), Very Low-density Lipoprotein (VLDL) (OR = 2.24), and Atherogenic Index of Plasma (AIP) (OR = 2.95) were independent risk factors for CSVD, while High-Density Lipoprotein (HDL) was a protective factor (OR = 0.33). RCS analysis revealed a J-shaped relationship between age and CSVD risk, and an inverted U-shaped relationship between VLDL and risk. Among the machine learning models, XGBoost achieved the best performance, with an AUC of 0.917 in the test set (95% CI: 0.866–0.969). SHAP analysis identified age as the most important global predictor and enabled visualization of individual patient-level risk-driving and protective factors. Conclusion: Aging, neutrophil-mediated inflammatory response, and an atherogenic lipid profile are independent risk factors for CSVD in patients with carotid plaque. Significant nonlinear effects were observed for age and VLDL. The XGBoost-based predictive model demonstrates excellent performance, and integration with the SHAP interpretability framework provides precise and intuitive decision support for early risk stratification and individualized intervention in CSVD.

Keywords: cerebral small vessel disease, carotid plaque, machine learning, SHAP, risk factors

1. Introduction

Cerebral Small Vessel Disease (CSVD) is a major cause of vascular cognitive impairment, gait disturbance, affective disorders, and stroke. Its pathological features are primarily characterized by white matter hyperintensities, lacunar infarction, cerebral microbleeds, and enlarged perivascular spaces [1, 2]. The incidence of CSVD is significantly higher in patients with carotid plaque, and carotid atherosclerosis is considered one of the key driving factors in the development and progression of CSVD [3]. However, in this patient population, the specific determinants that predominantly contribute to disease onset, as well as the underlying mechanisms through which these factors exert their effects, remain insufficiently clarified [4, 5]. Existing evidence suggests that chronic inflammatory responses, cerebral endothelial dysfunction, blood–brain barrier disruption, and their interactions constitute the principal pathogenic mechanisms of atherosclerosis-related CSVD [6]. In addition, modifiable risk factors such as hypertension and type 2 diabetes mellitus should not be overlooked. Although traditional statistical approaches can identify individual influencing factors, they are limited in their ability to elucidate complex nonlinear relationships and interactions among variables.

With the rapid advancement of artificial intelligence, machine learning has been increasingly applied in medical research. However, the limited interpretability of conventional machine learning models poses significant challenges for their clinical translation [7, 8]. The emergence of Explainable Artificial Intelligence (XAI) provides a promising solution to this problem, as it enables the direct quantification and visualization of each variable's contribution to predictive outcomes while maintaining high predictive accuracy [9, 10].

Based on the above background, this study employs a multimethod framework, integrating LASSO regression, multivariate logistic regression, Restricted Cubic Spline (RCS) analysis, and the SHAP interpretability framework. By combining traditional statistical methods with explainable machine learning techniques, we aim to comprehensively identify the key determinants of CSVD in patients with carotid plaque, elucidate their mechanisms and patterns of influence, and provide a scientific basis for early risk stratification and individualized therapeutic strategies.

2. Materials and methods

2.1. Study population

This retrospective case–control study consecutively enrolled patients diagnosed with carotid atherosclerotic plaques between January 2020 and December 2023 at a single center, based on carotid ultrasonography. Inclusion criteria were as follows: (1) presence of atherosclerotic plaque in the common carotid artery or internal carotid artery confirmed by carotid ultrasonography, defined as an intima–media thickness ≥1.5 mm; (2) completion of standardized brain Magnetic Resonance Imaging (MRI); and (3) availability of complete key laboratory data. Exclusion criteria included: (1) high-risk cardioembolic conditions such as atrial fibrillation or sick sinus syndrome; (2) moderate-to-severe intracranial arterial stenosis (> 50%) confirmed by MRA or CTA; (3) history of carotid endarterectomy or carotid stent implantation; (4) active large-vessel vasculitis; (5) complete carotid artery occlusion; (6) poor-quality imaging data insufficient for diagnosis; and (7) non-vascular neurological diseases such as brain tumors or demyelinating disorders. A total of 373 patients were ultimately included and classified into a CSVD-positive group (n = 207) and a CSVD-negative group (n = 166) according to the *Neuroimaging Standards for Cerebral Small Vessel Disease* [11]. The study protocol was approved by the Ethics Committee of Youjiang Medical University for Nationalities (Approval No. 202506001). Given the retrospective nature of the study, the requirement for informed consent was waived.

2.2. Clinical data collection

Baseline clinical data were extracted from the hospital electronic medical record system, including demographic characteristics (age and sex) and laboratory parameters. All laboratory measurements were obtained from peripheral venous blood samples collected within 24 hours of admission, including hypersensitive C-Reactive Protein (hs-CRP). Based on these primary laboratory indicators, several derived indices were calculated, including the atherogenic index of plasma (AIP = $\log_{10}$[TG/HDL]), CRP-to-HDL ratio, LDL-to-HDL ratio, Platelet-to-Lymphocyte Ratio (PLR), Monocyte-to-Lymphocyte Ratio (MLR), and Systemic Immune-inflammation Index (SII).

2.3. Imaging assessment

All patients underwent 3.0T Magnetic Resonance Imaging (MRI), including T1-weighted sequences. CSVD imaging features were evaluated strictly according to the *Neuroimaging Standards for Cerebral Small Vessel Disease*. Two experienced neuroradiologists, blinded to clinical data, independently assessed imaging findings. CSVD-related imaging markers included: (1) lacunar infarcts, defined as cerebrospinal fluid–like lesions ≤ 20 mm in diameter located in the subcortical or deep gray matter regions; (2) White Matter Hyperintensities (WMHs), defined as non-lacunar hyperintense lesions on T2-FLAIR sequences, with severity graded using the Fazekas scale; (3) cerebral microbleeds, defined as 2–5 mm focal hypointensities on Susceptibility-Weighted Imaging (SWI); and (4) enlarged perivascular spaces, defined as cerebrospinal fluid–like signal spaces along the course of perforating arteries, with a diameter ≤ 3 mm. In cases of disagreement between the two radiologists, a third senior neuroradiologist was consulted to reach a final consensus diagnosis.

2.4. Restricted Cubic Spline (RCS) analysis

This study employed a stepwise screening approach to identify continuous variables potentially associated with nonlinear effects. After adjustment for confounding factors, Restricted Cubic Spline (RCS) regression was used to examine exposure–response relationships between selected variables and CSVD risk. RCS allows continuous variables to be incorporated into regression models through smooth spline functions, avoiding information loss caused by conventional linear assumptions or arbitrary categorization (e.g., tertiles or quartiles), thereby enabling a more accurate representation of complex dose–response relationships. In this study, variables exhibiting potential nonlinearity were included in the RCS framework. Three knots were used by default, positioned at the 10th, 50th, and 90th percentiles based on data distribution. The number of knots was further optimized using the Bayesian Information Criterion (BIC) to achieve an optimal balance between model fit and parsimony.

2.5. Machine learning modeling and analysis

A machine learning-based framework was developed for CSVD risk prediction. Missing values in continuous variables were imputed using the median imputation method. To eliminate scale differences among features, all variables were standardized using Z-score normalization. Four machine learning algorithms were applied: a radial basis function kernel Support Vector Machine (SVM) with a fixed Cost parameter (C = 1.0), a K-Nearest Neighbors (KNN) classifier using Euclidean distance (k = 5), Partial Least Squares Discriminant Analysis (PLS-DA) with two latent components, and a gradient boosting decision tree (XGBoost) with a learning rate of 0.1 and a maximum tree depth of 3. The dataset was randomly split into a training set and a testing set at a 7:3 ratio. Model performance was evaluated using 10-fold cross-validation, with accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity as primary evaluation metrics. To enhance interpretability, SHapley Additive exPlanations (SHAP) were applied to quantify the

contribution of each feature to model predictions. Force plots and feature importance summary plots were generated to visualize decision mechanisms underlying key predictors. All analyses were performed in the R programming environment using relevant machine learning packages.

2.6. Statistical analysis

Statistical analyses were performed using SPSS version 22.0 and R version 4.0.3. Continuous variables with a normal distribution were expressed as mean ± standard deviation ($\overline{x} \pm s$) and compared using the independent samples t-test. Non-normally distributed continuous variables were expressed as median (interquartile range) and compared using the Mann–Whitney U test. Categorical variables were expressed as frequencies (percentages) and compared using the chi-square test χ^2. Multivariate logistic regression analysis was used to identify independent risk factors for CSVD. Restricted Cubic Spline (RCS) analysis was applied to explore potential nonlinear associations between continuous variables and CSVD risk. Receiver Operating Characteristic (ROC) curves were constructed to evaluate the predictive performance of the XGBoost model for CSVD, and the area under the curve (AUC) was calculated.

3. Results

3.1. Comparison of baseline clinical characteristics

A total of 373 patients with carotid atherosclerotic plaques were included in this study, comprising 126 females (33.8%) and 247 males (66.2%). Baseline comparisons showed that female patients were significantly older than male patients (64.64 ± 10.33 years vs. 58.84 ± 10.98 years, $p < 0.01$). In addition, Platelet count (PLT) was significantly higher in females than in males (293.57 ± 72.88 vs. 272.26 ± 71.92, $p = 0.007$). No statistically significant differences were observed between the two groups in the prevalence of Cerebral Small Vessel Disease (CSVD) (60.3% in females vs. 44.9% in males, $p = 0.181$), nor in CRP levels, neutrophil count, lipid profile parameters (HDL, LDL, VLDL, TG, CHOL), HbA1c, or inflammatory-derived indices (CRP/HDL, LDL/HDL, PLR, MLR, AIP, SII) (all $p > 0.05$). Detailed results are presented in Table 1.

Table 1. Comparison of general information

	Gender		t(χ^2) value	*p* value
	Female (n = 126)	Male (n = 247)		
Age (x ± s, years)	64.64 ± 10.33	58.84 ± 10.98	4.922	< 0.01
Cerebral small vessel disease (positive/negative)	76/50	131/161	1.791	0.181
CRP(x ± s, mg/L)	22.79 ± 31.21	19.52 ± 28.05	1.024	0.306
N(x ± s,×10⁹/L)	5.01 ± 1.90	5.32 ± 2.23	-1.317	0.189
PLT(x ± s,×10⁹/L)	293.57 ± 72.88	272.26 ± 71.92	2.694	0.007
L(x ± s,×10⁹/L)	2.06 ± 0.82	2.15 ± 2.89	-0.365	0.716
HDL(x ± s, mmol/L)	1.28 ± 0.59	1.24 ± 0.51	0.682	0.495
LDL(x ± s, mmol/L)	2.91 ± 0.99	2.75 ± 0.87	1.581	0.115

Table 1. Continued

VLDL(x ± s, mmol/L)	1.02 ± 0.64	1.07 ± 0.61	-0.644	0.520
TG(x ± s, mmol/L)	1.61 ± 0.80	1.75 ± 1.09	-1.419	0.157
CHOL(x ± s, mmol/L)	4.73 ± 1.32	4.53 ± 1.11	1.549	0.122
HbA1C(x ± s,%)	7.67 ± 2.90	8.41 ± 7.48	-1.064	0.288
CRP/HDL	23.90 ± 45.28	18.90 ± 28.55	1.302	0.194
LDL/HDL	2.70 ± 1.36	2.55 ± 1.20	1.075	0.283
PLR	161.37 ± 69.00	163.08 ± 77.47	-0.209	0.834
MLR	2.86 ± 1.82	3.49 ± 3.45	-1.927	0.055
AIP	0.21 ± 0.77	0.26 ± 0.72	-0.653	0.514
SII	828.45 ± 529.18	926.95 ± 979.31	-1.053	0.293

3.2. Variable selection and analysis of independent risk factors

3.2.1. LASSO regression-based feature selection

To identify key predictors of CSVD in patients with carotid plaque and address multicollinearity, Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied for feature selection. The optimal regularization parameter (λ) was determined using 10-fold cross-validation. The cross-validation error curve (Figure 1A) demonstrated a typical U-shaped trend of model deviance with increasing log(λ). Based on the minimum error criterion, the optimal λ was selected at $\log(\lambda) \approx -6.3$, where the model achieved an optimal balance between predictive accuracy and generalizability. The coefficient profile plot (Figure 1B) illustrates the shrinkage process of predictor coefficients as regularization intensity increases. Ultimately, 13 variables with non-zero coefficients were retained (see Appendix File 1), including age, sex, CRP, Neutrophil count (N), Platelet count (PLT), Lymphocyte count (L), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Very Low-Density Lipoprotein (VLDL), glycated hemoglobin (HbA1c), and composite indices such as CRP/HDL, LDL/HDL, and the Atherogenic Index of Plasma (AIP). These variables collectively represent demographic characteristics, inflammatory status, and lipid metabolic profiles, forming a comprehensive feature set for subsequent analyses. This integrated indicator system not only captures basic patient information but also reflects systemic inflammatory responses and lipid metabolism processes, providing a robust foundation for downstream modeling and interpretation.

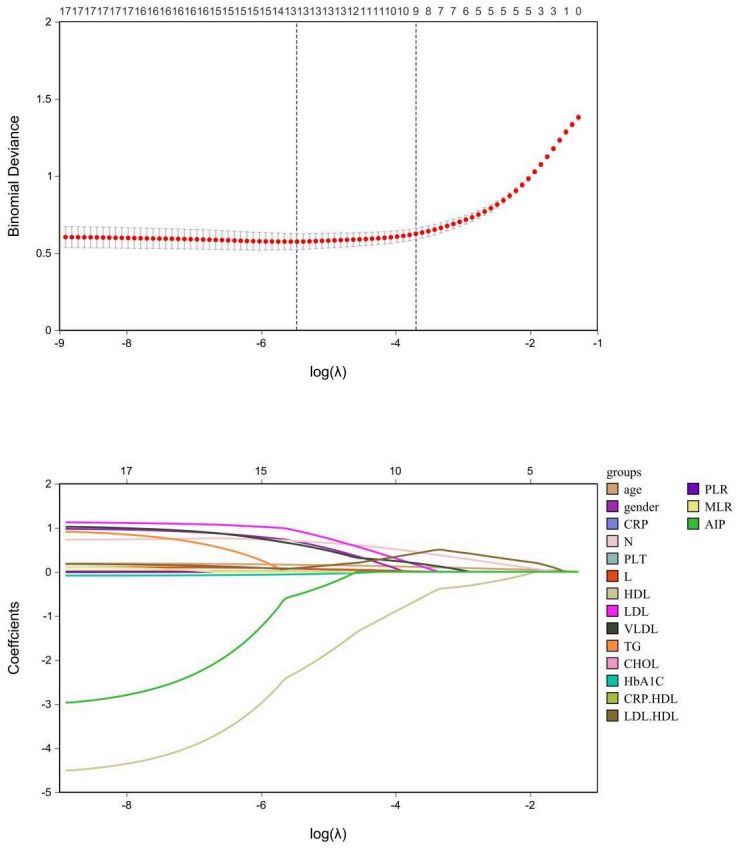

Figure 1. Results of LASSO regression feature selection. (A) Cross-validation error curve of LASSO regression; (B) coefficient paths of variables under LASSO regularization

3.2.2. Univariate and multivariate logistic regression analysis

Based on variables selected by LASSO regression, univariate logistic regression was first performed to preliminarily assess the associations between each factor and CSVD. The results indicated that age (OR = 1.12, 95% CI: 1.08–1.16), CRP (OR = 1.02, 95% CI: 1.01–1.03), Neutrophil count (N) (OR = 2.31, 95% CI: 1.89–2.82), PLT (OR = 1.01, 95% CI: 1.01–1.02), HDL (OR = 0.22, 95% CI: 0.12–0.41), LDL (OR = 2.05, 95% CI: 1.52–2.76), VLDL (OR = 3.17, 95% CI: 1.86–5.41), and AIP (OR = 4.29, 95% CI: 2.45–7.52) were significantly associated with CSVD risk (all $p < 0.01$).

In multivariate logistic regression analysis, after adjustment for potential confounders, age (OR = 1.09, 95% CI: 1.04–1.14, $p < 0.01$), Neutrophil count (N) (OR = 1.87, 95% CI: 1.42–2.46, $p < 0.01$), HDL (OR = 0.33, 95% CI: 0.16–0.68, $p = 0.003$), LDL (OR = 1.78, 95% CI: 1.22–2.60, $p = 0.003$), VLDL (OR = 2.24, 95% CI: 1.18–4.25, $p = 0.013$), and AIP (OR = 2.95, 95% CI: 1.58–5.52, $p < 0.01$) remained independent predictors of CSVD. These findings suggest that aging, neutrophil-mediated inflammatory activation, atherogenic lipid profiles (low HDL and elevated LDL/VLDL), and dysregulated lipid metabolism (elevated AIP) are independent risk factors for CSVD in patients with carotid plaque (Table 2).

Table 2. Univariate and multivariate logistic regression analysis

Variables	Univariate					Multivariate				
	β	S.E	Z	*p*	OR (95%CI)	β	S.E	Z	*p*	OR (95%CI)
Age	0.15	0.02	9.12	<0.01	1.16 (1.12 ~ 1.20)	0.19	0.03	6.87	<0.01	1.20 (1.14 ~ 1.27)
CRP(mg/L)	0.01	0.00	2.95	0.003	1.01 (1.01 ~ 1.02)	0.01	0.02	0.40	0.686	1.01 (0.96 ~ 1.06)
N(×10⁹/L)	0.74	0.09	8.09	<0.01	2.10 (1.76 ~ 2.52)	0.82	0.15	5.29	<0.01	2.26 (1.67 ~ 3.07)
PLT(×10⁹/L)	0.01	0.00	7.12	<0.01	1.01 (1.01 ~ 1.02)	0.02	0.00	4.45	<0.01	1.02 (1.01 ~ 1.02)
L(×10⁹/L)	0.27	0.14	1.90	0.057	1.31 (0.99 ~ 1.74)					
HDL(mmol/L)	-2.09	0.29	-7.24	<0.01	0.12 (0.07 ~ 0.22)	-3.41	1.18	-2.90	0.004	0.03 (0.00 ~ 0.33)
LDL(mmol/L)	0.77	0.14	5.62	<0.01	2.16 (1.65 ~ 2.83)	1.38	0.59	2.35	0.019	3.99 (1.26 ~ 12.66)
VLDL(mmol/L)	0.95	0.20	4.73	<0.01	2.57 (1.74 ~ 3.81)	0.96	0.40	2.43	0.015	2.62 (1.20 ~ 5.71)
HbA1C(%)	-0.00	0.02	-0.05	0.959	1.00 (0.97 ~ 1.03)					
CRP/HDL	0.02	0.00	4.43	<0.01	1.02 (1.01 ~ 1.03)	0.02	0.03	0.59	0.553	1.02 (0.96 ~ 1.07)
LDL/HDL	1.08	0.13	8.36	<0.01	2.95 (2.29 ~ 3.81)	-0.09	0.58	-0.15	0.883	0.92 (0.29 ~ 2.87)
AIP	0.92	0.16	5.68	<0.01	2.50 (1.82 ~ 3.43)	-1.09	0.51	-2.13	0.033	0.34 (0.12 ~ 0.92)
Gender										
0					1.00 (Reference)					
1	-0.30	0.22	-1.34	0.181	0.74 (0.48 ~ 1.15)					

3.3. Restricted cubic spline analysis

Restricted Cubic Spline (RCS) analysis was performed to assess dose–response relationships between the selected variables and CSVD risk. The results showed that age was significantly associated with CSVD risk in a nonlinear manner (p for overall < 0.001, p for nonlinear = 0.023), with a J-shaped association. The risk increased slowly before the age of 70 years, but rose sharply thereafter, reaching nearly 100% in individuals over 90 years of age. Similarly, Very Low-Density Lipoprotein (VLDL) exhibited a significant nonlinear association with CSVD risk (p for overall < 0.001, p for nonlinear < 0.001; where p for overall indicates the overall significance of the association and p for nonlinear indicates the significance of the nonlinear component). The relationship followed an inverted U-shaped curve, with the risk peaking at approximately 1.5 mmol/L VLDL, as estimated by identifying the point where the derivative of the fitted spline function equals zero (Figure 2A–B).

The remaining variables demonstrated significant linear associations with CSVD risk (all p for overall < 0.001), with no evidence of nonlinear effects (AIP: p for nonlinear = 0.227; HDL: p for nonlinear = 0.803; LDL: p for nonlinear = 0.121; Neutrophil count [N]: p for nonlinear = 0.155; Platelet count [PLT]: p for nonlinear = 0.138; all p for nonlinear > 0.05). Specifically, AIP, LDL, neutrophil count, and platelet count were positively associated with CSVD risk, whereas HDL was negatively associated with risk (protective factor) (Figure 2C–G). Overall, these findings indicate that, except for age and VLDL, which exhibit threshold and nonlinear effects, most lipid- and inflammation-related biomarkers show stable linear dose–response relationships with CSVD risk.

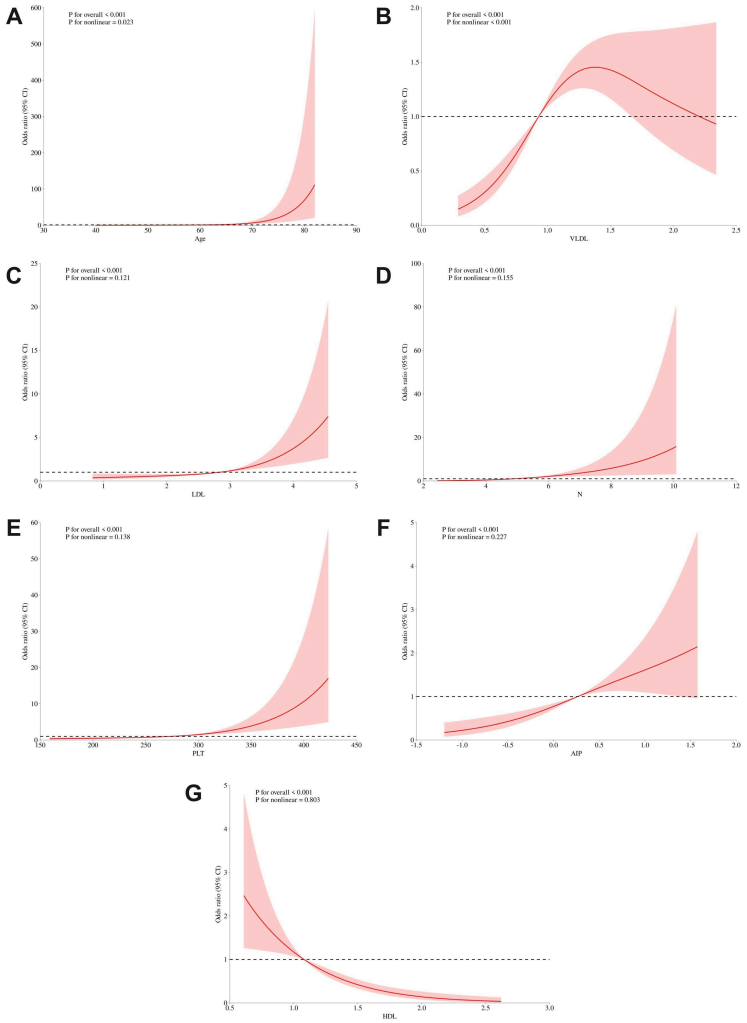

Figure 2. Restricted cubic spline curves for key indicators and CSVD risk. (A) Age; (B) VLDL; (C) LDL; (D) Neutrophil count; (E) Platelet count; (F) Plasma Atherogenic Index; (G) HDL

3.4. Construction and evaluation of four machine learning models

Based on multivariate logistic regression analysis of the training and testing datasets, seven statistically significant features were selected, including age, platelet count, neutrophil count, HDL, LDL, VLDL, and the Plasma Atherogenic Index (AIP). Using these variables, four machine learning algorithms were developed, and their performance was evaluated using five-fold cross-validation. Receiver Operating Characteristic (ROC) curve analysis demonstrated that the XGBoost model achieved superior predictive performance for CSVD in patients with carotid plaque, with an AUC of 0.917 (95% CI: 0.866–0.969), significantly outperforming the other models (Table 3, Figures 3A–B). Furthermore, Decision Curve Analysis (DCA) indicated that the XGBoost model provided a consistently higher net clinical benefit across most threshold probability ranges in both the training and validation sets (Figures 3C–D). Additional performance curves and AUC forest plots are provided in Appendix Files 2–5. Based on these results, the XGBoost model was selected for subsequent SHAP-based interpretation and visualization.

Table 3. AUC values of four machine learning models in the training and validation sets

	Model	AUC	AUC_CI
Training set	SVM_Kernel	0.958846348456738	0.959 (0.939-0.978)
	NeighborMethod	0.844942935852027	0.845 (0.798-0.891)
	PLSModel	0.944004047900152	0.944 (0.920-0.968)
	XGBoost	0.955529319165683	0.956 (0.935-0.976)
Validation set	SVM_Kernel	0.907037037037037	0.907 (0.850-0.964)
	NeighborMethod	0.819259259259259	0.819 (0.740-0.899)
	PLSModel	0.891111111111111	0.891 (0.832-0.950)
	XGBoost	0.917407407407407	0.917 (0.866-0.969)

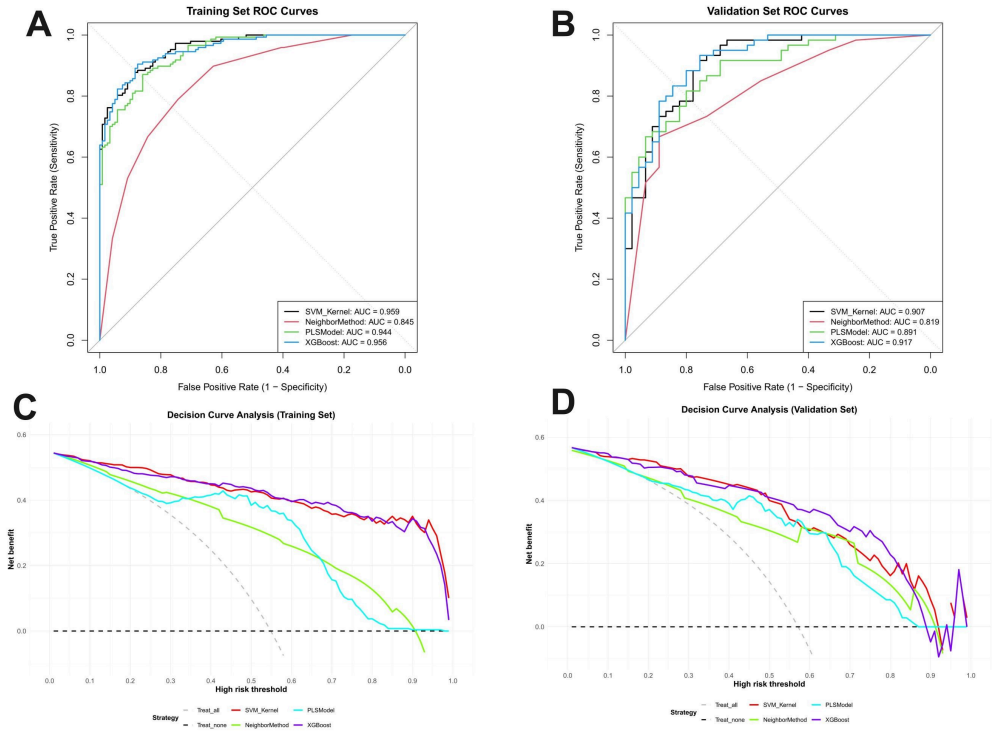

Figure 3. ROC and DCA curves of four machine learning models. (A–B) ROC curves for training and validation sets; (C–D) Decision Curve analysis (DCA) curves

3.5. SHAP-based interpretation of the XGBoost model

3.5.1. Global interpretation

Based on the SHapley Additive exPlanations (SHAP) framework, the decision-making process of the XGBoost model was interpreted from a global perspective. The results demonstrated that among all features, age exhibited the highest predictive importance, with a mean absolute SHAP value of 0.187, far exceeding that of the other variables. This dominant effect was consistently observed across SHAP analyses of XGBoost, random forest, and neural network models. Specifically, the SHAP value for Neutrophil count (N) was 0.124, while that for Platelet count (PLT) was 0.081, indicating relatively weaker contributions to model predictions.

Notably, the Plasma Atherogenic Index (AIP) showed the lowest contribution among all features, with a SHAP value of only 0.006 (Figure 4A).

The SHAP summary plot (beeswarm plot; Figure 4B) further clarified the directional relationships between features and CSVD risk. Increased values of age, neutrophil count, platelet count, LDL, and VLDL were associated with higher positive SHAP values, indicating an increased predicted risk of CSVD. In contrast, higher HDL levels were associated with negative SHAP values, suggesting a protective effect. The distribution width of SHAP values indicated the magnitude of each variable's impact on model output, with age exhibiting the widest range, further confirming its central role in prediction.

Analysis of the SHAP dependence plots (Figure 4C) revealed pronounced nonlinear relationships between key variables and model output. From the perspective of age, SHAP values increased progressively with advancing age, with a particularly pronounced upward trend observed after 70 years, consistent with the findings from the RCS analysis. For inflammatory markers, Neutrophil count (N) showed a positive association with CSVD risk, with a more evident risk increase when N exceeded 6 × 10⁹/L. Regarding lipid-related indicators, HDL was negatively associated with risk, whereas LDL and VLDL were positively associated. Notably, VLDL exhibited an inverted U-shaped relationship with CSVD risk, with the highest risk observed at intermediate concentrations.

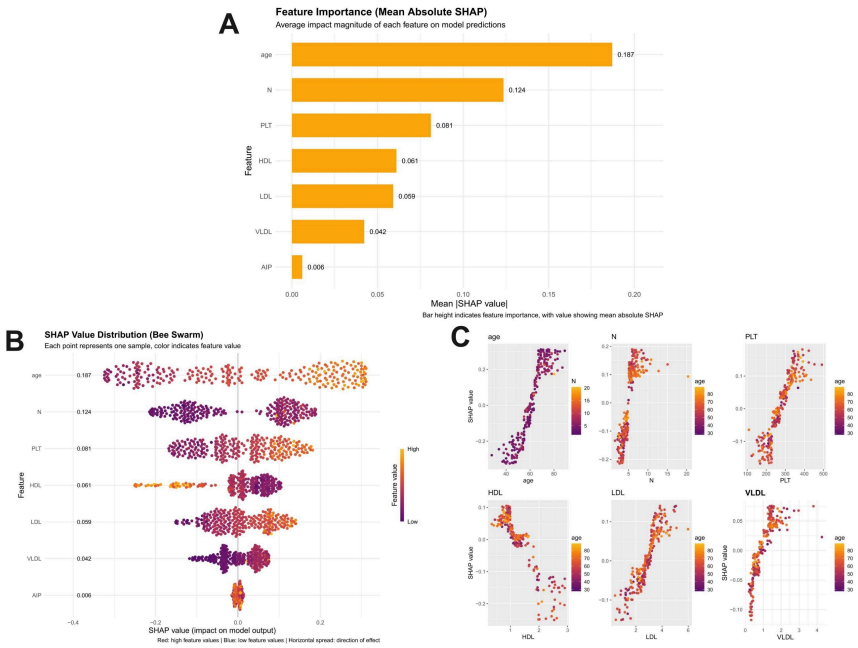

Figure 4. Global SHAP interpretation. (A) Feature importance (mean absolute SHAP values); (B) SHAP summary (beeswarm) plot; (C) SHAP dependence plots

3.5.2. Local interpretation

Individual-level risk assessment based on the XGBoost model demonstrated strong predictive performance across typical case categories, including high-risk, low-risk, and intermediate-to-high-risk patients (Figure 5A–F). Additional SHAP-based local explanations for other cases are provided in Appendix File 6. In terms of risk stratification, distinct patterns of feature contributions were observed across groups. In the high-risk group, dyslipidemia—characterized by elevated LDL and reduced HDL—was the dominant driver of risk, although Neutrophil count (N) and age showed relatively protective or attenuating effects in certain cases. In the low-risk group, protection was primarily associated with younger age, low inflammatory burden, and

favorable lipid profiles. In the intermediate-to-high-risk group, a complex interaction between risk and protective factors was observed, where the adverse effects of older age and elevated LDL were counterbalanced by protective influences such as higher HDL and lower Platelet count (PLT). These findings not only clarify the interaction mechanisms among key variables such as age, N, LDL, and HDL in individualized prediction, but also provide a visual interpretability tool for clinical decision-making, supporting precise risk stratification and management of CSVD.

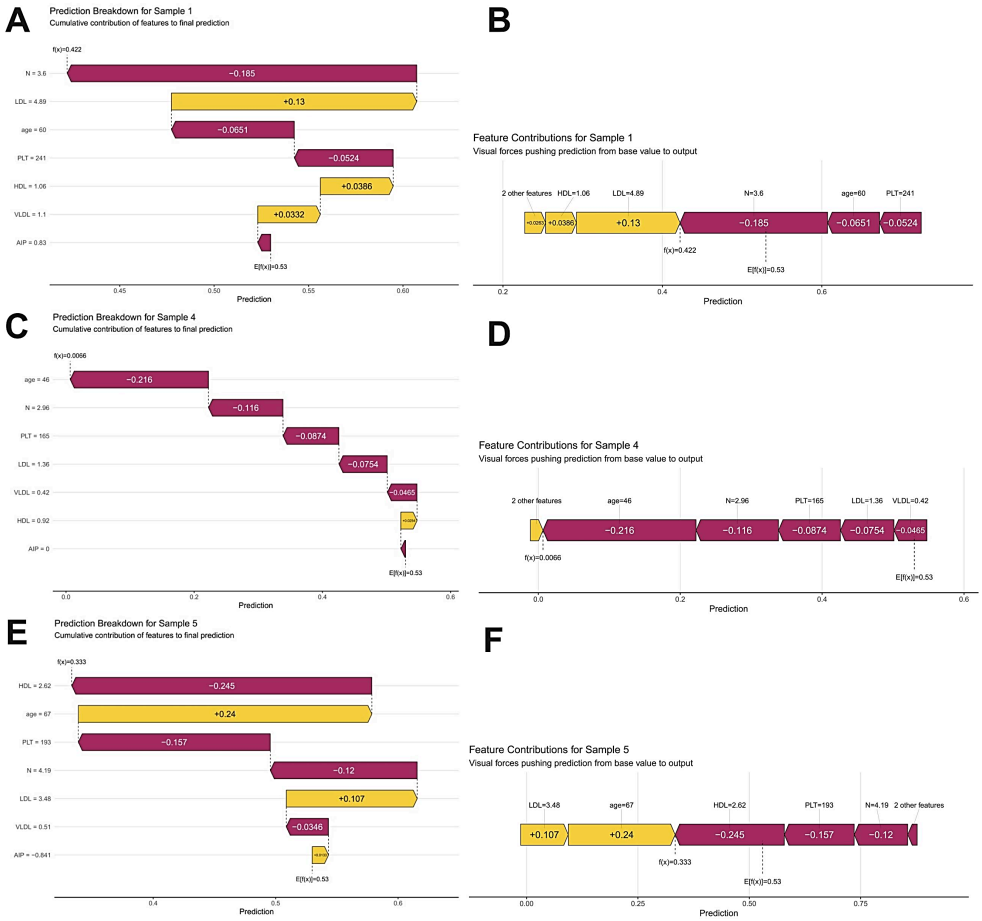

Figure 5. SHAP-based local interpretation of CSVD risk in representative samples. (A) Raw visualization of Sample 1; (B) feature contribution plot of Sample 1; (C) cumulative feature contribution plot of Sample 4; (D) raw visualization of Sample 4; (E) cumulative feature contribution plot of Sample 5; (F) feature contribution plot of Sample 5

4. Discussion

Through an integrated analysis of clinical and imaging data from 373 patients with carotid atherosclerotic plaques, this study combined traditional statistical methods with explainable machine learning approaches to systematically investigate the influencing factors and underlying mechanisms of Cerebral Small Vessel Disease (CSVD). The results demonstrated that age, neutrophil count, and lipid-related indicators (including high-density lipoprotein, low-density lipoprotein, and very low-density lipoprotein), as well as the Plasma

Atherogenic Index (AIP), were independent risk factors for CSVD [12, 13]. Among these, age and VLDL exhibited significant nonlinear associations with CSVD risk. In comparative model analysis, the XGBoost model achieved the best predictive performance, with an Area Under the Curve (AUC) of 0.917, and further enabled individualized risk interpretation through the SHAP framework.

The findings indicate that CSVD risk is closely associated with aging. Restricted Cubic Spline (RCS) analysis showed that the risk of CSVD increased markedly after 70 years of age, which may be attributed to age-related structural alterations in cerebral microvasculature, including reduced vascular elasticity and disruption of the blood–brain barrier. Regarding inflammatory markers, neutrophil count demonstrated independent predictive value beyond conventional inflammatory indicators such as C-Reactive Protein (CRP), suggesting that neutrophil-mediated immune responses may play a direct role in microcirculatory injury in CSVD. From the perspective of lipid metabolism, reduced HDL and elevated LDL and VLDL were significantly associated with increased CSVD risk, potentially through mechanisms involving atherosclerosis progression and lipotoxicity. Notably, an interesting finding emerged in the lipid-related analyses: the direction of the Odds Ratio (OR) for the Plasma Atherogenic Index (AIP) in multivariate logistic regression was inconsistent with that observed in univariate analysis. This discrepancy may be attributed to multicollinearity within the model. As a composite indicator derived from Triglycerides (TG) and HDL, AIP may share overlapping information with directly included lipid variables such as VLDL, HDL, and LDL, leading to instability in its estimated independent effect. This suggests that when interpreting highly correlated variables, a comprehensive evaluation using multiple models—including interpretable machine learning outputs such as SHAP values—is necessary. The relatively low SHAP contribution of AIP further supports this interpretation.

RCS analysis further revealed nonlinear associations of age and VLDL with CSVD risk. Age demonstrated a J-shaped relationship with CSVD, with 70 years identified as a critical inflection point. This finding highlights the importance of prioritizing screening and preventive interventions for CSVD in carotid plaque patients aged over 70 years. VLDL exhibited an inverted U-shaped relationship with CSVD risk, with peak risk observed at moderate concentrations. This may be explained by the stronger pro-inflammatory and endothelial-damaging effects of VLDL within this intermediate range, whereas at higher concentrations, other metabolic abnormalities may dominate the risk profile. These findings provide evidence for more precise lipid management strategies in clinical practice [14, 15].

Among the evaluated machine learning models, XGBoost demonstrated consistently superior performance across all metrics. Compared with traditional methods such as sUpport Vector Machines (SVM), K-Nearest Neighbors (KNN), and Partial Least Squares Discriminant Analysis (PLS-DA), XGBoost offers advantages in parallel computation, handling of missing data, and predictive accuracy. Decision Curve Analysis (DCA) further confirmed that the model provided substantial net clinical benefit across a wide range of threshold probabilities, supporting its potential utility in the auxiliary diagnosis of CSVD. The integration of SHAP interpretation effectively addressed the "black-box" limitation of machine learning models. From a global perspective, key predictors such as age, Neutrophil count (N), and Platelet count (PLT) were identified, while SHAP summary and dependence plots revealed complex interactions and nonlinear effects among variables. At the individual level, SHAP-based explanations enabled patient-specific risk decomposition, demonstrating that dyslipidemia plays a dominant role in high-risk patients, whereas High-Density Lipoprotein (HDL) exerts a protective effect. These interpretable outputs provide an intuitive visualization tool to support precise risk stratification and personalized management strategies for CSVD.

This study has several limitations. First, its single-center retrospective design inevitably introduces selection bias, which may limit the generalizability of the findings. Second, the sample size is relatively limited, particularly in subgroups with markedly elevated VLDL levels, which may affect the robustness of

nonlinear association estimates. Third, although model performance was evaluated using internal validation, external validation using independent datasets is essential to ensure generalizability and clinical applicability, as emphasized in studies of predictive model validation.

To address these limitations, future research should focus on three directions: conducting multicenter prospective cohort studies with larger sample sizes and integrating multi-omics biomarkers such as cytokine profiles and metabolomics to enhance robustness and translational value; combining deep learning approaches with radiomics features to further improve CSVD predictive accuracy and clinical applicability; and validating the SHAP-based interpretability framework through clinical intervention studies to facilitate the translation of research findings into individualized therapeutic strategies.

5. Conclusion

To investigate the mechanisms underlying Cerebral Small Vessel Disease (CSVD) in patients with carotid atherosclerotic plaques, this study innovatively integrated classical statistical approaches with explainable machine learning algorithms. The findings indicate that three major factors jointly contribute to the pathogenesis of CSVD: patient age, systemic inflammatory status, and dysregulated lipid metabolism. Notably, specific indicators exhibited pronounced nonlinear relationships. The predictive model developed based on the XGBoost algorithm demonstrated excellent discriminative performance, providing a solid foundation for potential clinical application. Furthermore, by incorporating the SHAP interpretability framework, the model offers transparent and interpretable risk stratification, thereby providing robust scientific decision support for early screening, risk assessment, and the development of individualized therapeutic strategies for CSVD.

References

[1] Xu, Y., Song, Y., Tang, T., Jia, W., Xu, H., Li, Y., Guo, Y., Wang, X., & Liu, R. (2025). Correlation between neuroimaging scores and carotid artery ultrasound features in cerebral small vessel disease patients. *Cerebrovascular Diseases Extra, 15*, 93. https: //doi.org/10.1159/000000000

[2] Litak, J., Mazurek, M., Kulesza, B., Szmygin, P., Litak, J., Kamieniak, P., & Grochowski, C. (2020). Cerebral small vessel disease. *International Journal of Molecular Sciences, 21*(24), 9729. https: //doi.org/10.3390/ijms21249729

[3] Lv, M., Yang, X., Shi, X., Cao, S., Li, W., Zhou, M., Gou, X., & Huang, Y. (2025). Daqinjiao decoction ameliorates CSV via RXR-γ/PPAR-γ/VEGF-α pathway: Insights from transcriptome sequencing and network pharmacology. *Journal of Cellular and Molecular Medicine, 29*(8), e70712. https: //doi.org/10.1111/jcmm.70712

[4] Stoisavljevic, S., Zdraljevic, M., Radojicic, A., Pavlovic, A., & Mijajlovic, M. (2024). Carotid artery stenosis is related to cerebral small vessel disease magnetic resonance imaging burden. *Heliyon, 10*(7), e36052. https: //doi.org/10.1016/j.heliyon.2024.e36052

[5] Elahi, F. M., Wang, M. M., & Meschia, J. F. (2023). Cerebral small vessel disease-related dementia: More questions than answers. *Stroke, 54*(2), 648–657. https: //doi.org/10.1161/STROKEAHA.122.040867

[6] Zhai, F., Yang, M., Wei, Y., Wang, M., Gui, Y., Han, F., Zhou, L., Ni, J., Yao, M., & Zhang, S. (2020). Carotid atherosclerosis, dilation, and stiffness relate to cerebral small vessel disease. *Neurology, 94*(16), e1811–e1820. https: //doi.org/10.1212/WNL.0000000000009183

[7] Li, B., Eisenberg, N., Beaton, D., Lee, D. S., Al-Omran, L., Wijeysundera, D. N., Hussain, M. A., Rotstein, O. D., de Mestral, C., & Mamdani, M. (2024). Using machine learning to predict outcomes following

transfemoral carotid artery stenting. *Journal of the American Heart Association, 13*(8), e035425. https: //doi.org/10.1161/JAHA.124.035425

[8] Eini, P., Eini, P., Serpoush, H., Rezayee, M., & Tremblay, J. (2025). Machine learning models for carotid artery plaque detection: A systematic review of ultrasound-based diagnostic performance. *Journal of Stroke and Cerebrovascular Diseases, 34*(2), 108446. https: //doi.org/10.1016/j.jstrokecerebrovasdis.2024.108446

[9] Guo, T., Wang, M., Wang, C., Gao, D., Liu, Y., Geng, Z., Fan, M., Zhu, H., Chen, L., & Qiu, B. (2025). Development and validation of an interpretable machine learning model for cerebral small vessel disease risk assessment. *International Journal of Medical Informatics, 204*, 106070. https: //doi.org/10.1016/j.ijmedinf.2025.106070

[10] Wang, Y., Li, Y., Jiao, S., Pan, Y., Deng, X., Qin, Y., Zhao, D., & Liu, Z. (2024). Correlation analysis and predictive model construction of metabolic syndrome, complete blood count-derived inflammatory markers, and overall burden of cerebral small vessel disease. *Heliyon, 10*(14), e35065. https: //doi.org/10.1016/j.heliyon.2024.e35065

[11] Duering, M., Biessels, G. J., Brodtmann, A., Chen, C., Cordonnier, C., de Leeuw, F. E., Debette, S., Frayne, R., Jouvent, E., Rost, N. S., & Wardlaw, J. M. (2023). Neuroimaging standards for research into small vessel disease—Advances since 2013. *The Lancet Neurology, 22*(7), 602–613. https: //doi.org/10.1016/S1474-4422(23)00131-2

[12] Dupré, N., Drieu, A., & Joutel, A. (2024). Pathophysiology of cerebral small vessel disease: A journey through recent discoveries. *The Journal of Clinical Investigation, 134*(8), e172841. https: //doi.org/10.1172/JCI172841

[13] Inoue, Y., Shue, F., Bu, G., & Kanekiyo, T. (2023). Pathophysiology and probable etiology of cerebral small vessel disease in vascular dementia and Alzheimer's disease. *Molecular Neurodegeneration, 18*(1), 46. https: //doi.org/10.1186/s13024-023-00654-8

[14] Yang, S., & Webb, A. J. S. (2023). Associations between neurovascular coupling and cerebral small vessel disease: A systematic review and meta-analysis. *European Stroke Journal, 8*(5), 895–906. https: //doi.org/10.1177/23969873231189425

[15] Yuan, H., Zhu, B., Li, C., & Zhao, Z. (2023). Ceramide in cerebrovascular diseases. *Frontiers in Cellular Neuroscience, 17*, 1191609. https: //doi.org/10.3389/fncel.2023.1191609

Appendix

Table A1. Specific features selected by LASSO regression

(Intercept)	-17.45842408
age	0.161202899
gender	0.68620876
CRP	0.011484966
N	0.699453451
PLT	0.012555752
L	0.071628178
HDL	-2.285948682
LDL	0.936557452
VLDL	0.62171102
HbA1C	-0.055379948
CRP/HDL	0.006984849

LDL/HDL	0.079966982
AIP	-0.523015082

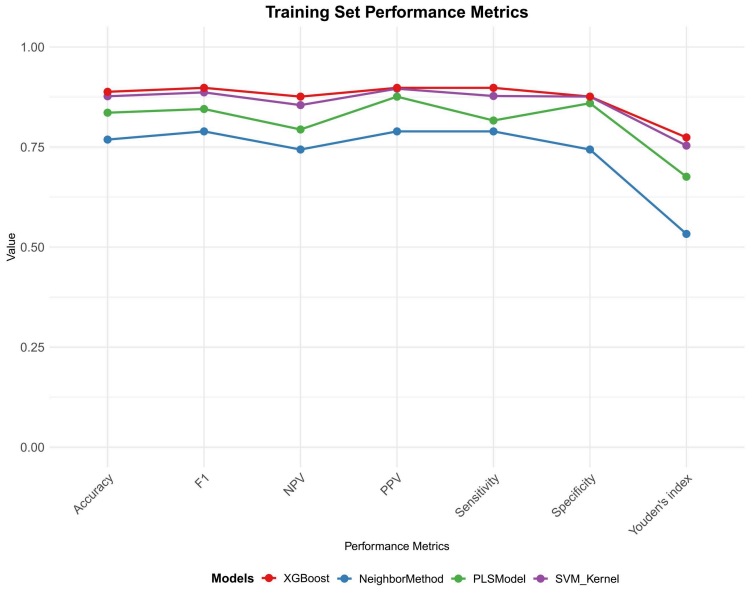

Figure A1. Performance curves in the training set

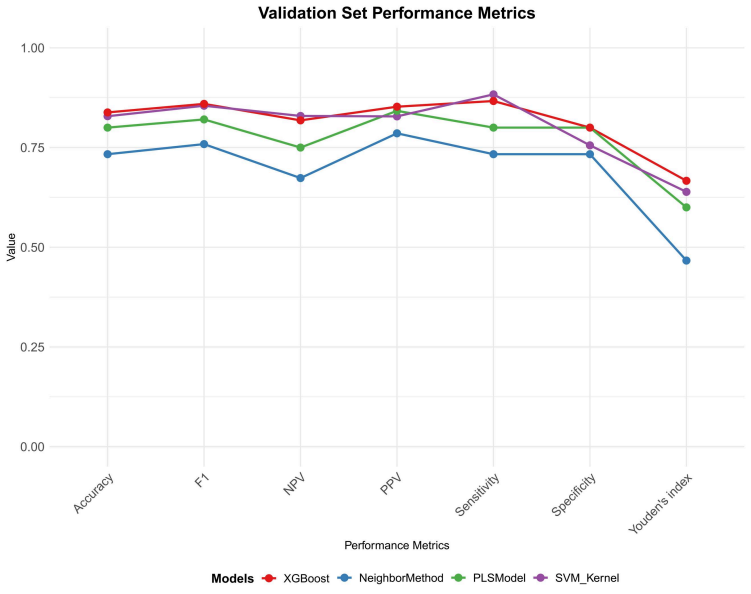

Figure A2. Performance curves in the validation set

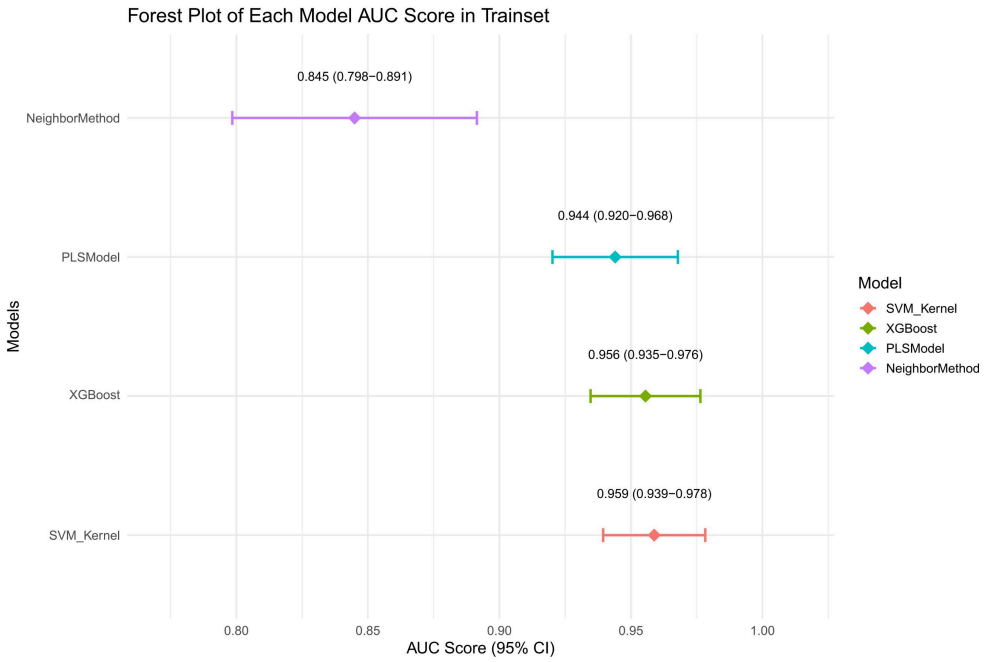

Figure A3. AUC forest plot in the training set

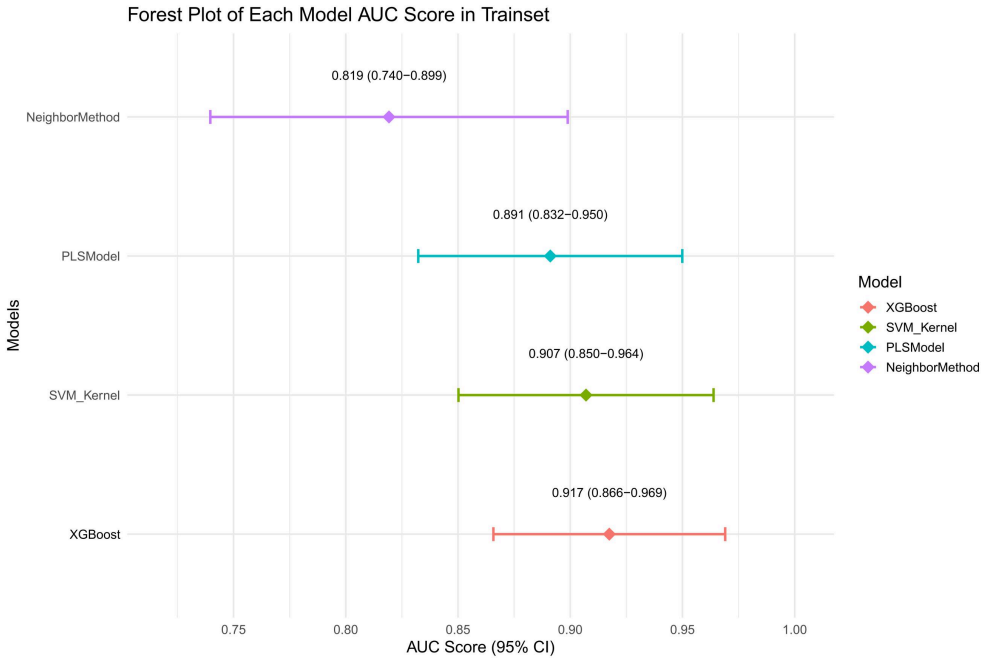

Figure A4. AUC forest plot in the validation set

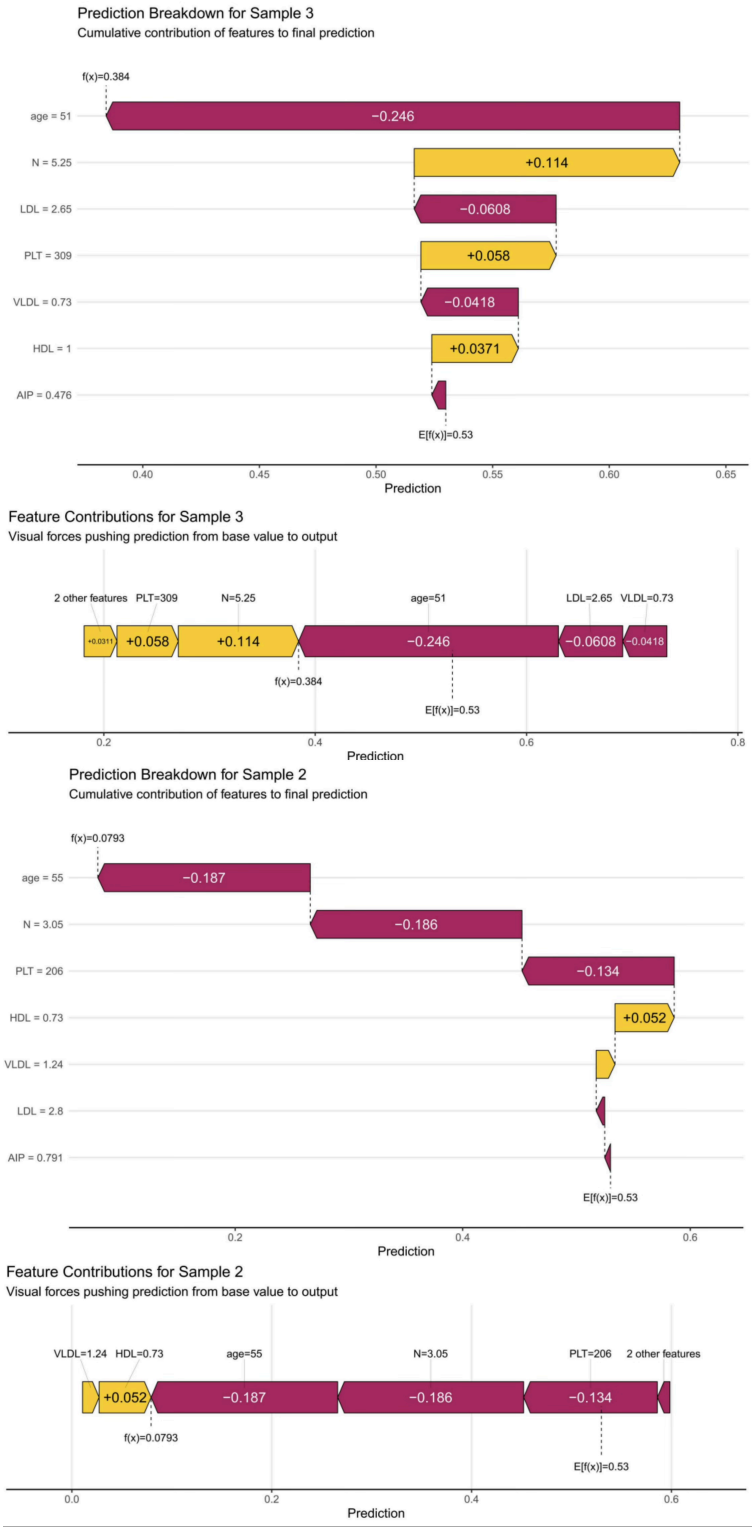

Figure A5. SHAP-based local interpretations for additional cases