TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 15/24 pp1495–1508
DOI: 10.26599/TST.2023.9010092
Volume 29, Number 5, October 2024

Machine Learning for Selecting Important Clinical Markers of Imaging Subgroups of Cerebral Small Vessel Disease Based on a Common Data Model

Lan Lan, Guoliang Hu, Rui Li, Tingting Wang, Lingling Jiang, Jiawei Luo, Zhiwei Ji, and Yilong Wang*

Abstract: Differences in the imaging subgroups of cerebral small vessel disease (CSVD) need to be further explored. First, we use propensity score matching to obtain balanced datasets. Then random forest (RF) is adopted to classify the subgroups compared with support vector machine (SVM) and extreme gradient boosting (XGBoost), and to select the features. The top 10 important features are included in the stepwise logistic regression, and the odds ratio (OR) and 95% confidence interval (CI) are obtained. There are 41 290 adult inpatient records diagnosed with CSVD. Accuracy and area under curve (AUC) of RF are close to 0.7, which performs best in classification compared to SVM and XGBoost. OR and 95% CI of hematocrit for white matter lesions (WMLs), lacunes, microbleeds, atrophy, and enlarged perivascular space (EPVS) are 0.9875 (0.9857–0.9893), 0.9728 (0.9705–0.9752), 0.9782 (0.9740–0.9824), 1.0093 (1.0081–1.0106), and 0.9716 (0.9597–0.9832). OR and 95% CI of red cell distribution width for WMLs, lacunes, atrophy, and EPVS are 0.9600 (0.9538–0.9662), 0.9630 (0.9559–0.9702), 1.0751 (1.0686–1.0817), and 0.9304 (0.8864–0.9755). OR and 95% CI of platelet distribution width for WMLs, lacunes, and microbleeds are 1.1796 (1.1636–1.1958), 1.1663 (1.1476–1.1853), and 1.0416 (1.0152–1.0687). This study proposes a new analytical framework to select important clinical markers for CSVD with machine learning based on a common data model, which has low cost, fast speed, large sample size, and continuous data sources.

Key words: common data model; machine learning; cerebral small vessel disease; imaging subgroups; clinical markers

1 Introduction

The clinical symptoms of cerebral small vessel disease (CSVD) are insidious and the course of the disease is slow. It is often called “little stroke” and involves the catastrophic damage of small blood vessels in the

• Lan Lan and Rui Li are with IT Center, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China. E-mail: lanlan@bjtth.org; lirui@bjtth.org.
• Guoliang Hu, Tingting Wang, and Yilong Wang are with Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China. E-mail: guoliang_hu@163.com; wangtingtingdr@sina.com; yilong528@aliyun.com.
• Lingling Jiang is with China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China. E-mail: linglingjiang712@aliyun.com.
• Jiawei Luo is with West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610044, China. E-mail: luojiawei@wchscu.cn.
• Zhiwei Ji is with College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210095, China. E-mail: Zhiwei.Ji@njau.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-12-12; revised: 2023-06-10; accepted: 2023-07-05

whole brain. The decline of cognitive function is the most common and important clinical manifestation of CSVD. About half of vascular cognitive impairment is caused by CSVD. At present, the diagnosis of CSVD is mainly based on magnetic resonance imaging (MRI), including several types of imaging features: white matter lesions (WMLs), lacunes, microbleeds, atrophy, enlarged perivascular space (EPVS), etc. The pathogenesis of CSVD has been studied[1−5].

Ryu et al.[6] concluded that alkaline phosphatase (ALP) was associated with white matter hyperintensities (WMHs) and cerebral infarction, but not with cerebral microbleeds by using quantitative and logistic regression (LR) with 1082 neurologically healthy subjects. Lee et al.[7] found that ALP was related to CSVD by using LR and generalized additive model with 1011 neurologically healthy participants. Piao et al.[8] found that ALP was related to lacunes and WMHs by using LR with 568 participants. Wada et al.[9] found that the association between C-reactive protein and small vessel disease related brain lesions was not significant by using LR with 689 individuals. Mitaki et al.[10] found that the high sensitive C-reactive protein was related to lacunes by using LR with 519 neurologically normal subjects. Hassan et al.[11] found that hyperhomocysteinaemia was a risk factor for CSVD by using LR with 172 patients and 172 controls. Nam et al.[12] obtained a dose-dependent relationship between total homocysteine and CSVD by using difference tests with 1578 participants. Cao et al.[13] obtained total homocysteine and lacunes correlations by using generalized linear model and LR with 1023 participants. Wada et al.[14] obtained the correlation between uric albumin and CSVD by using LR with 651 individuals. Chung et al.[15] obtained that 25 hydroxyvitamin D and lacunes, WMHs, and cerebral microbleeds were both related by using linear regression with 838 patients. Park et al.[16] obtained the effect of hemoglobin on atrophy by using LR with 2040 participants. Vilar-Bergua et al.[17] found that the urinary albumin to creativity ratio was related to CSVD by using LR with 1037 subjects. Kim et al.[18] found that total bilirubin was not related to CSVD by using LR with 1128 subjects. Yin et al.[19] obtained high density lipoprotein cholesterol and apolipoprotein A-1 that were associated with WMLs in women by using LR with 848 subjects. Nam et al.[20] found that triglyceride glucose was related to CSVD by using linear regression with 2615 neurologically healthy participants. Kang et al.[21] found that insulin like growth factor-1 was associated with cognitive function in CSVD patients by using a partial correlation analysis with 216 patients. Chu et al.[22] obtained the correlation between subclinical hydropyroidism and CSVD by using LR with 354 individuals. Oberheiden et al.[23] used difference tests with 24 patients to evaluate the role of platelets and cellular coagulation activation in CSVD. Jiang et al.[4] found that the neutral count was related to the enlarged perivascular spaces and lacunes by using LR and generalized linear model. Karel et al.[24] used LR and random forest (RF) with 80 CSVD patients and 38 health individuals to evaluate the relationship between biomarkers from blood samples and CSVD.

Due to the high cost of MRI, the sample size of studies on laboratory markers and CSVD ranges from hundreds to thousands. Most of them select specific laboratory markers first, and then use LR to study the relationship between laboratory markers and CSVD. These studies have reported the differences between the characteristics of patients with CSVD and those without CSVD, but the differences in the imaging subgroups of patients diagnosed with CSVD need to be further explored.

In order to select important clinical markers of imaging subgroups, this study first extracts a large number of data through information system of a hospital to build a special database of CSVD according to the observational medical outcomes partnership, common data model (OMOP-CDM) standard[25], which ensures a large sample size, low cost, and continuous data sources. Then, by using machine learning method after propensity score matching (PSM), important laboratory markers for imaging subgroups of CSVD are selected based on the data. Finally, the important markers are put into stepwise LR to obtain odds ratio (OR) and 95% confidence interval (CI). This study provides a new analytical framework for identifying important laboratory markers for each imaging subgroups of CSVD, which are important for subsequent clinical validation.

2 Method

2.1 Data source

Data comes from several information systems of Beijing Tiantan Hospital, Capital Medical University

from January 1, 2012 to February 8, 2021, including hospital information system (HIS), electronic medical record (EMR), laboratory information management system (LIS), and picture archiving and communication system (PACS).

2.2 Study patients

For the purpose of including adult inpatients diagnosed as CSVD, we consider two aspects of inclusion criteria of study patients: diagnosis and head MRI imaging report. (1) Retrieving from diagnostic name using CSVD, small vascular disease, and hereditary CSVD or retrieving from the diagnosis code (ICD-10) using I63.801, E85.400x027, E85.414, I65.800x008, I72.003, I67.800x005, I67.800x012, E75.205, and G71.300x001. (2) Retrieving from head MRI imaging report using white matter hyperintensity (WMH), leukoencephalopathy, leukoaraiosis, ischemic leukoaraiosis, patchy white matter, lacunes, microbleeds, Fazekas, atrophy and senile brain changes, EPVS, CSVD, and small vascular disease. The adult inpatients who meet one of the above retrieval conditions are included in this study. Non angiogenic WMLs such as multiple sclerosis, white matter dysplasia, and metabolic encephalopathy retrieved from diagnosis name or head MRI imaging report are excluded from this study. The imaging report is processed on the basis of the regular expression. Then, according to the construction standard of OMOP-CDM[25], we build the OMOP-CSVD database.

2.3 Outcomes

Considering the availability of data, five imaging subgroups of CSVD are used as outcomes of this study including WMLs, lacunes, microbleeds, atrophy, and EPVS. If any words such as leukoencephalopathy, white matter disease, leukoaraiosis, ischemic leukoaraiosis, patchy white matter, and Fazekas are retrieved from head MRI imaging report, the patient is marked with WML. If any lacunas other than new lacunar infarction are retrieved from head MRI imaging report or lacunar cerebral infarction retrieved from disease history, the patient is marked as lacunar. If microbleed appears in head MRI imaging report, the patient is marked with cerebral microbleed. If atrophy or senile brain changes appear in head MRI imaging report, the patient is marked with brain atrophy. If EPVS appears in head MRI imaging report, the patient is marked with EPVS.

In the patients' head MRI reports, the physicians describe the symptom characteristics in standard, short Chinese words. Then, we divide the whole patients into 5 symptom subgroups by using regular expression. Next we evaluate the results of subgroups. 10% of the patients in each group are randomly selected. Experienced physicians review record of every patient in the sample. If the accuracy rate is greater than 90%, the subgroups are considered acceptable, otherwise all patients in the subgroups would be checked one by one, and then the sampling would continue. At last, the accuracy of 5 subgroups are all more than 90%, and we consider our subgrouping acceptable.

2.4 Features

Laboratory markers with high test frequency such as blood routine test, urine routine test, and biochemical test are included in this study. The values of laboratory markers measured for the first time after admission are adopted. Demographic characteristics (age and sex), risk factors (smoking and drinking), and disease history (hypertension, diabetes, hyperlipidemia, heart disease, and stroke) are also included in this study.

2.5 Data preprocessing

Markers with a certain missing rate may have systematic data missing issues, which may have a significant impact on model training and classification[26, 27]. Therefore, we first delete laboratory markers with a missing rate greater than 30%. Then, we use the winsorizing method[28] to process outliers. In this study, we consider that values less than 1% or more than 99% of quantiles are outliers. Outliers less than 1% are replaced by random numbers between 1% and 5% of quantiles, and outliers more than 99% are replaced by random numbers between 95% and 99% of quantiles. After that, multiple imputation is used to fill in the missing data. Multiple imputation[29] is a method of processing missing values based on repeated simulation. It generates a complete set of data from a dataset containing missing values. The missing data in each dataset is filled with Monte Carlo method.

2.6 Feature selection

PSM[30] is used to obtain a balanced positive and negative samples (Fig. 1a). First, age and gender are identified as confounding variables that affect imaging subgroups of CSVD. Second, the propensity scores are estimated. Third, matched samples using the propensity scores are created. After matching, we use Gini index

of RF to rank features importance[31, 32]. Gini index evaluates that how much each feature contributes to each tree in RF, and then takes an average value. Then, the top 10 of 65 features are input to stepwise LR. After stepwise regression, the features retained in the model are both important and not seriously multicollinearity.

2.7 Statistical analysis

2.7.1 Classification model

RF[33] made up of a collection of decision trees is used to classify the five outcomes (binary variable). By inputting features into the RF model, outcomes are identified. We also use two other machine learning models, support vector machine (SVM)[34], and extreme gradient boosting (XGBoost)[35], to classify image subgroups and compare their performance with RF.

2.7.2 Experimental setup

This study use randomForest, e1071, and XGBoost packages in R (Version 4.2.1) to implement RF, SVM, and XGBoost, respectively. Hyperparameter is not adjustable in the traditional sense, but should be set high enough[36]. So we set the number of decision trees to 1000 (default is 500 in R). Bernard et al.[37] suggested that for classification tasks, the number of variables used for binary tree in the node should be set as the square root of the total number of variables to obtain best performance. Therefore, we set the parameter as the quadratic root of the number of variables in the dataset (default in R). A decision tree with a minimum number of nodes (node size) of 1 (default in R) can provide good results[36, 38]. So we set the node size to 1. We use the default parameter settings in R when training SVM and XGBoost.

Since the parameters have been determined, the data is divided into training set and test set according to the ratio of 7∶3, and repeated the modeling for 30 times. The average performance of test set is taken, and the classification evaluation matrix (accuracy, precision, recall, F1 score, and area under curve (AUC)). The importance of features to outcomes is output through the best classification model. Histograms are used to analyze the distribution of important markers.

2.7.3 Statistical model

In order to explain important features, we use LR[39] to model the top 10 important features, and output OR and 95% CI (Fig. 1b). All analysis are performed using R.

3 Result

3.1 Patients characteristics

There are 37 558 adult inpatients and 41 290 adult inpatient records diagnosed with CSVD included in this study before matching. The situation of hospitalization records (N=41 290) is analyzed. The average age of these patients is 67.19±12.45 years and

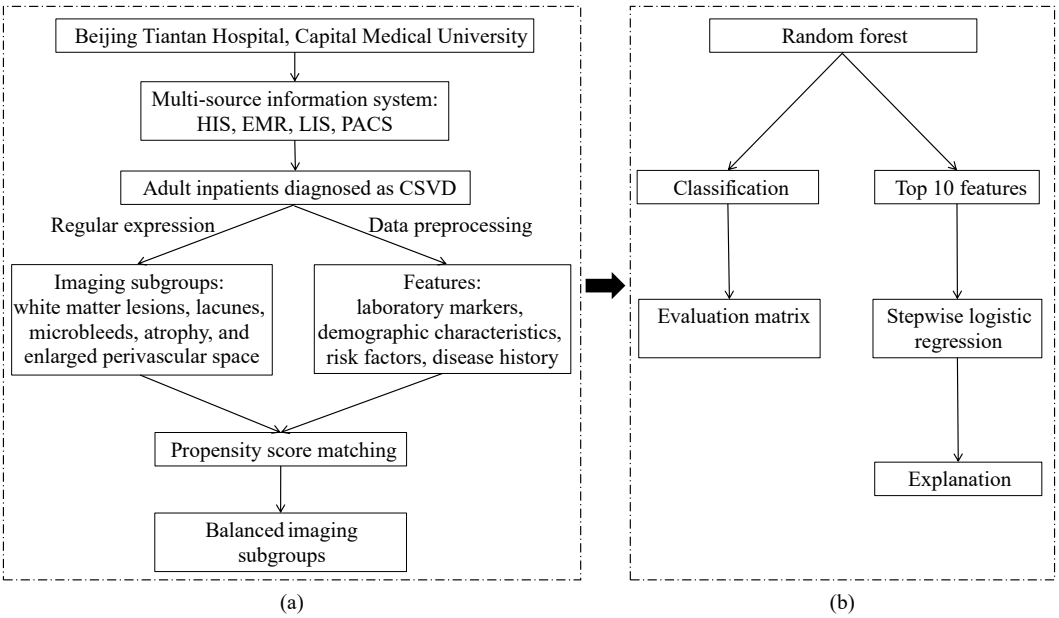


Fig. 1 Framework of this study.

62.03% are males. The numbers of patients for risk factors, history of disease, and imaging subgroups of CSVD are shown in Table 1.

3.2 Classification performance of imaging subgroups using machine learning

After matching, the number of positive and negative imaging subgroups is the same. The balanced dataset is then used for classification. Table 2 shows that the accuracy, precision and AUC of SVM, and XGBoost are lower than those of RF. The classification accuracy of WMLs, lacunes, and microbleeds is almost 0.7000, atrophy's accuracy is over 0.6000, and accuracy of EPVS is close to 0.6000 with RF.

We further analyze the distribution characteristics and differences of top 3 important markers between correctly classified and incorrectly classified patients. Figure 2 shows that patients with high platelet distribution width (PDW) are more likely to be correctly classified for WMLs. Figure 3 shows that for lacunes classification, patients with low hematocrit are more likely to be correctly classified. Figure 4 shows that patients with low hematocrit are more likely to be correctly classified for microbleeds. Figure 5 shows that the distribution of the top three important markers in the two groups of patients is similar for atrophy. Figure 6 shows that patients with low hematocrit are more likely to be correctly classified for EPVS.

Table 1 Demographic characteristics, habits, and diseases of patients.

Characteristic	Characteristic	*N*	Percentage (%)
Demographic characteristic	Gender, male	25 614	62.03
Risk factor	Smoking	15 641	37.88
	Drinking	12 454	30.16
History of disease	Hypertension	22 665	54.89
	Diabetes	9679	23.44
	Hyperlipidemia	5016	12.15
	Heart disease	888	2.15
	Stroke	6637	16.07
Imaging subgroup	WMLs	23 650	57.28
	Lacunes	12 932	31.32
	Microbleeds	3378	8.18
	Atrophy	19 615	47.51
	EPVS	302	0.73

Table 2 Classification performance of imaging subgroups using machine learning models.

Machine learning model	Imaging subgroup	Accuracy	Precision	Recall	F1	AUC
RF	WMLs	0.6872	0.7136	0.6265	0.6672	0.6873
	Lacunes	0.6898	0.7480	0.5738	0.6494	0.6900
	Microbleeds	0.6992	0.7461	0.6094	0.6706	0.6998
	Atrophy	0.6399	0.6427	0.6272	0.6348	0.6399
	EPVS	0.5843	0.5930	0.5854	0.5861	0.5865
SVM	WMLs	0.6777	0.7056	0.6109	0.6547	0.6777
	Lacunes	0.6797	0.7339	0.5641	0.6376	0.6797
	Microbleeds	0.6873	0.7212	0.6133	0.6623	0.6875
	Atrophy	0.6353	0.6377	0.6262	0.6318	0.6354
	EPVS	0.5732	0.5839	0.5844	0.5687	0.5765
XGBoost	WMLs	0.6739	0.6798	0.6589	0.6691	0.6739
	Lacunes	0.6798	0.7053	0.6173	0.6584	0.6798
	Microbleeds	0.6777	0.6838	0.6595	0.6712	0.6778
	Atrophy	0.6280	0.6281	0.6278	0.6279	0.6280
	EPVS	0.5483	0.5422	0.5491	0.5430	0.5502

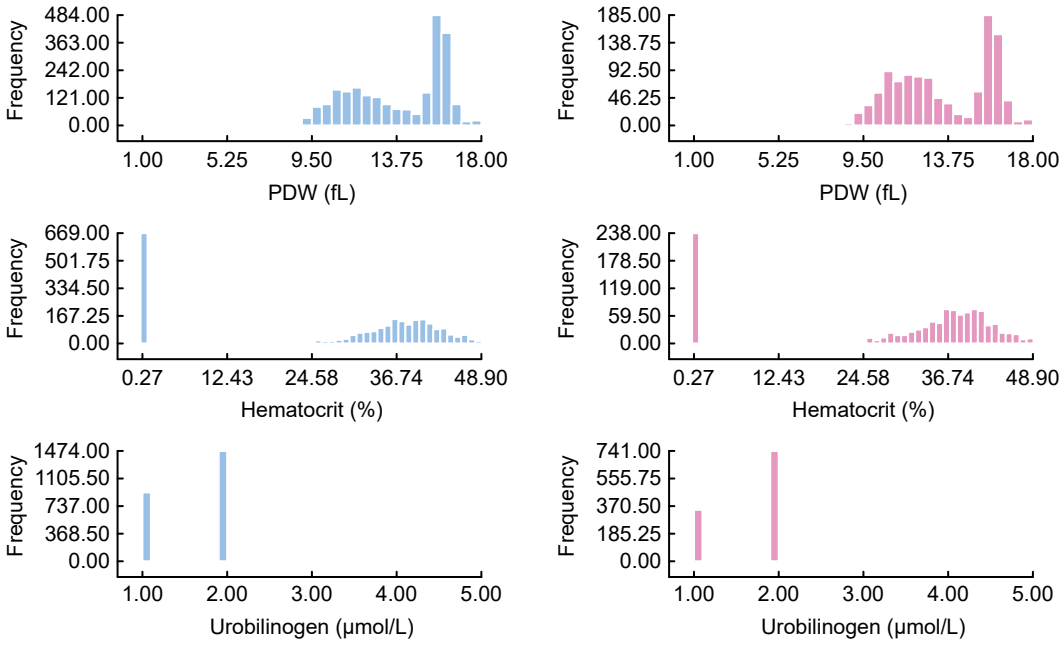


Fig. 2 Histograms of the top three important markers in the population with correct and incorrect classification for white matter lesions (left: correct classification; right: incorrect classification).

3.3 Top 10 important markers of imaging subgroups

Table 3 shows that the top 10 important markers obtained from RF are all laboratory markers. There are differences and similarities in important markers of each imaging subgroups of CSVD. Hematocrit is the top 10 markers for the five imaging subgroups. PDW is the top 10 markers for WMLs, lacunes, microbleeds, and EPVS. The red cell distribution width (RDW) is the top 10 markers for WMLs, lacunes, atrophy, and EPVS. Urobilinogen is the top 10 markers for WMLs, lacunes, and microbleeds. Creatinine, cholinesterase, and platelets are the top 10 markers for WMLs and atrophy. Platelet large cell ratio (P-LCR) and platelet

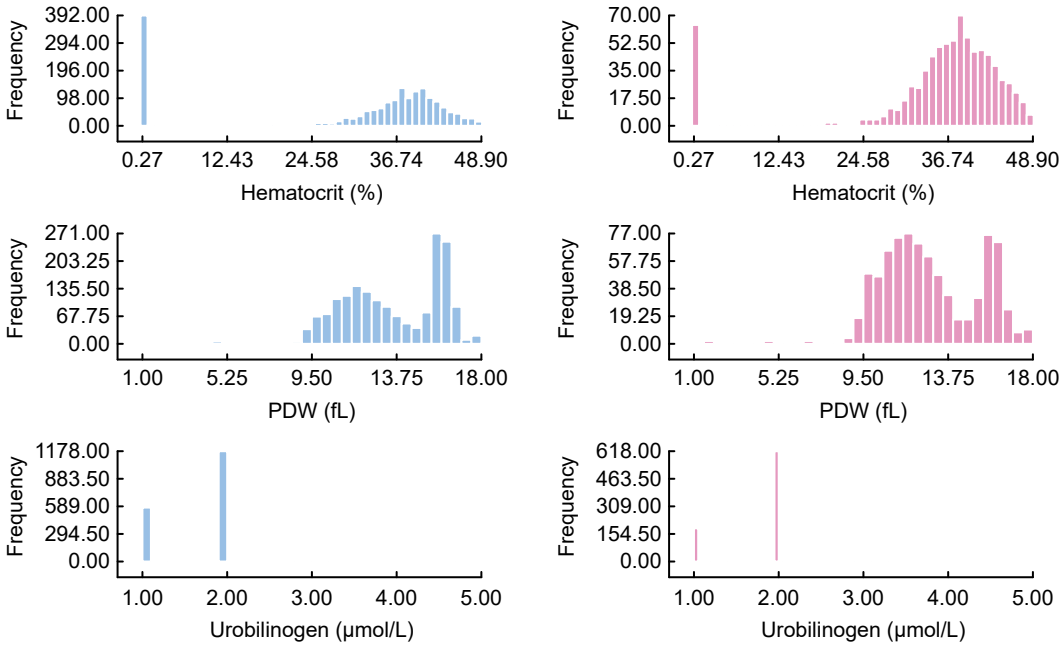


Fig. 3 Histograms of the top three important markers in the population with correct and incorrect classification for lacunes (left: correct classification; right: incorrect classification).

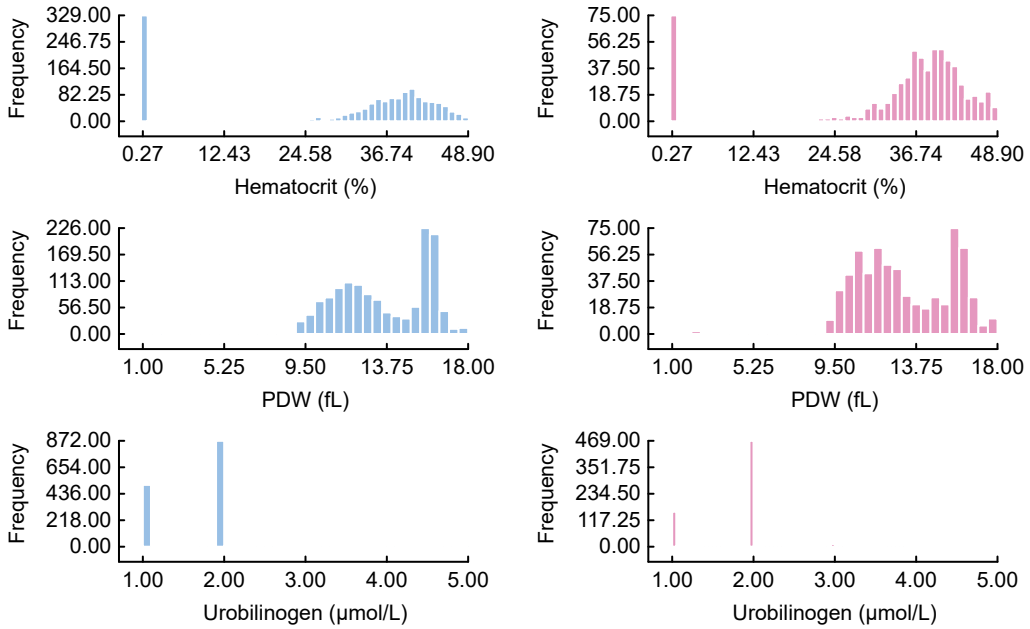


Fig. 4 Histograms of the top three important markers in the population with correct and incorrect classification for microbleeds (left: correct classification; right: incorrect classification).

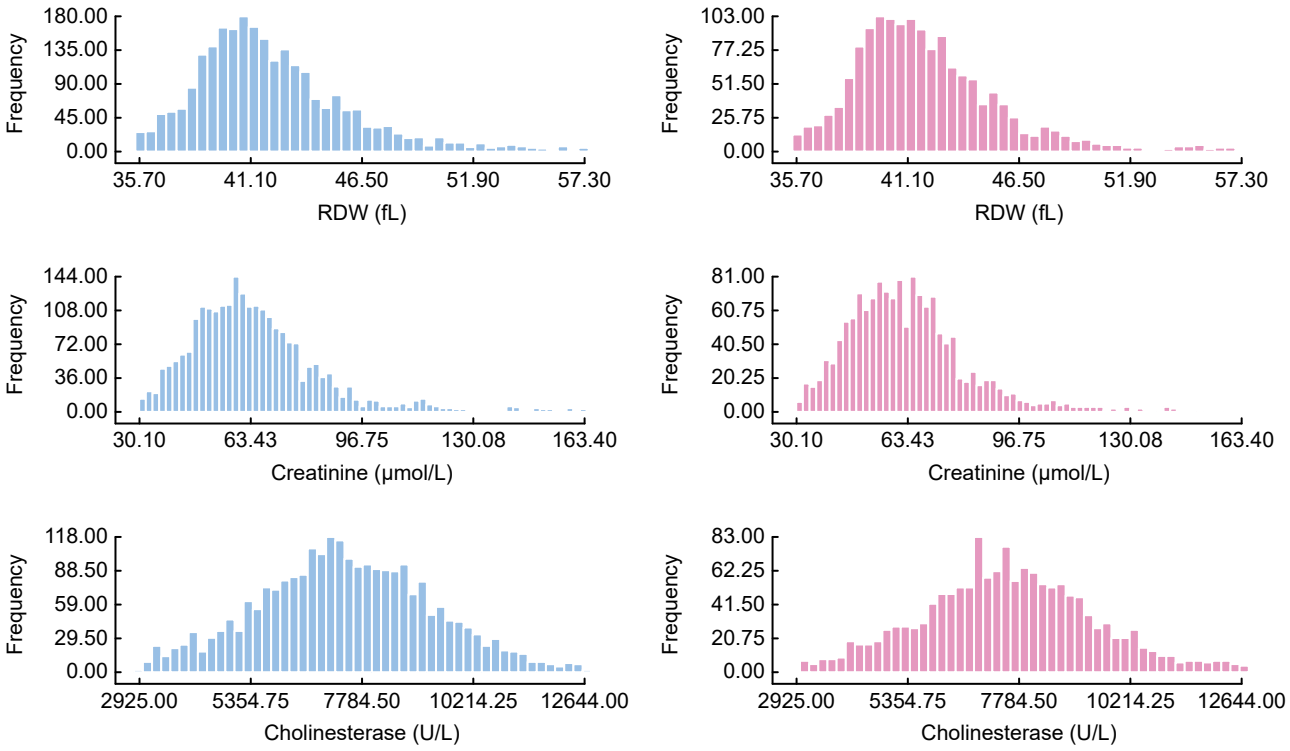


Fig. 5 Histograms of the top three important markers in the population with correct and incorrect classification for atrophy (left: correct classification; right: incorrect classification; RDW: red cell distribution width).

mean volume (PMV) are the top 10 markers for WMLs and lacunes. Alanine aminotransfer (ALT) is the top 10 markers for lacunes and atrophy. Glucose is the top 10 markers for lacunes and microbleeds. Phosphorus is the top 10 markers for lacunes and EPVS. Bilirubin and potassium are the top 10 markers for microbleeds and EPVS. Each imaging subgroup also has its own unique top 10 important markers.

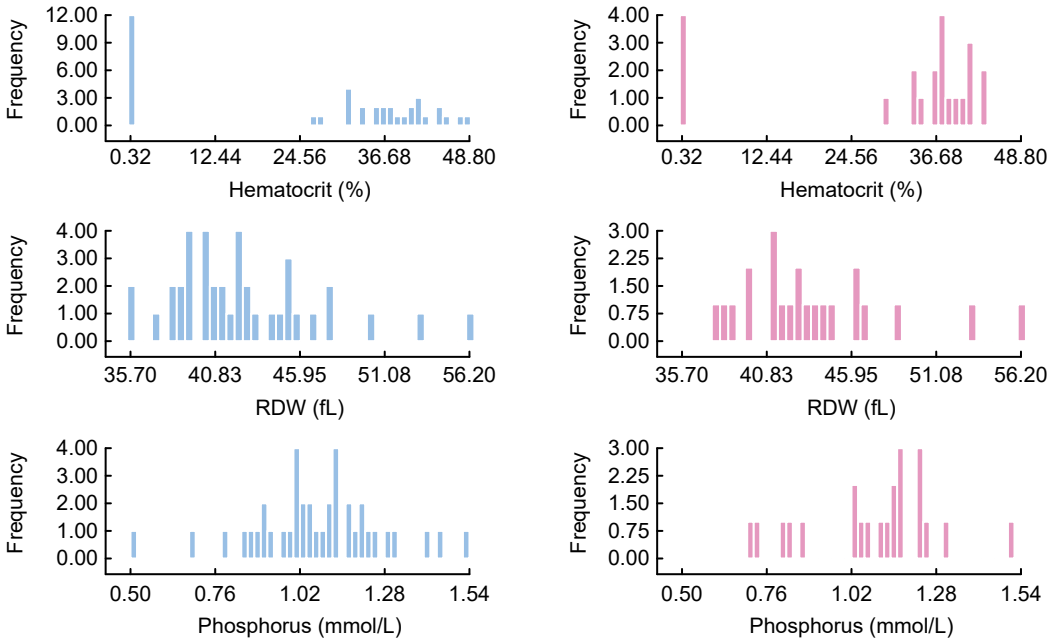


Fig. 6 Histograms of the top three important markers in the population with correct and incorrect classification for enlarged perivascular space (left: correct classification; right: incorrect classification).

Table 3 Top 10 important markers of five imaging subgroups from RF.

Number	WMLs	Lacunes	Microbleeds	Atrophy	EPVS
1	PDW	Hematocrit	Hematocrit	RDW	Hematocrit
2	Hematocrit	PDW	PDW	Creatinine	RDW
3	Urobilinogen	Urobilinogen	Urobilinogen	Cholinesterase	Phosphorus
4	Creatinine	P-LCR	Indirect bilirubin	ALT	Total bilirubin
5	RDW	RDW	Total bilirubin	Platelets	Direct bilirubin
6	Cholinesterase	Total cholesterol	Specific gravity of Urine	WBC	PDW
7	P-LCR	ALT	Glucose	Urate	LDH
8	Platelets	PMV	Eosinophils/100 leukocytes	BUN	Potassium
9	CO_2	Glucose	Potassium	Neutrophils	MCH
10	PMV	Phosphorus	ALP	Hematocrit	Indirect bilirubin

Note: P-LCR: platelet large cell ratio; PMV: platelet mean volume; ALT: alanine aminotransferase; ALP: alkaline phosphatase; WBC: white blood cell; BUN: urea nitrogen; LDH: lactate dehydrogenase; MCH: mean corpuscular hemoglobin.

3.4 Explanation for markers

The top 10 important markers are then input to stepwise LR to get OR and 95% CI. For WMLs, the 10 markers are retained in the model and have statistical significance. The OR of PDW, cholinesterase, platelets, CO_2, and PMV is greater than 1. The OR of hematocrit, creatinine, RDW, and P-LCR is less than 1. With negative as reference, the OR of urobilinogen is less than 1 (Table 4).

For lacunes, 9 of the 10 markers are retained in the model and have statistical significance. Markers with OR greater than 1 are PDW, PMV, and glucose. The OR of the remaining 6 markers is less than 1 (Table 5).

For microbleeds, the 10 markers are retained in the model, but one factor is not statistically significant. The OR of PDW, indirect bilirubin, eosinophils/100 leukocytes, and alkaline phosphatase is greater than 1. The OR of the remaining markers is less than 1 (Table 6).

For atrophy, the 10 markers are retained in the model, but one factor is not statistically significant. Markers with OR greater than or less than 1 account for half of the total (Table 7).

For EPVS, 5 of the 10 markers are retained in the model and have statistical significance. Except that the

Table 4 Regression results for WMLs.

Marker	OR	95% CI	
		Lower limit	Upper limit
PDW***	1.1796	1.1636	1.1958
Hematocrit***	0.9875	0.9857	0.9893
Urobilinogen	—	—	—
+-***	0.5994	0.5653	0.6356
1+*	0.7137	0.5484	0.9280
2+	1.1108	0.6227	2.0207
3+	0.9362	0.4893	1.8540
Creatinine***	0.9909	0.9897	0.9921
RDW***	0.9600	0.9538	0.9662
Cholinesterase***	1.0001	1.0000	1.0001
P-LCR***	0.9687	0.9635	0.9739
Platelet***	1.0017	1.0013	1.0020
CO_2***	1.0172	1.0099	1.0245
PMV**	1.0381	1.0094	1.0678

Note: ***P<0.001, **P<0.01, and *P<0.05.

Table 5 Regression results for lacunes.

Marker	OR	95% CI	
		Lower limit	Upper limit
Hematocrit***	0.9728	0.9705	0.9752
PDW***	1.1663	1.1476	1.1853
Urobilinogen	—	—	—
+-***	0.6028	0.5604	0.6484
1+*	0.7285	0.5419	0.9795
2+	0.9845	0.5176	1.9256
3+	1.5418	0.6514	4.2723
P-LCR***	0.9623	0.9561	0.9684
RDW***	0.9630	0.9559	0.9702
Total cholesterol***	0.9189	0.8953	0.9431
PMV**	1.0540	1.0196	1.0899
Glucose**	1.0181	1.0056	1.0307
Phosphorus***	0.4971	0.4371	0.5652

Note: ***P<0.001, **P<0.01, and *P<0.05.

OR of total bilirubin is greater than 1, the OR of other markers is less than 1 (Table 8).

4 Discussion

Based on a common data model with large sample size, low cost, and continuous data sources, this study selects the markers that are important to the phenotype of CSVD imaging from a large number of laboratory markers through machine learning models. This study adopts data driven feature selection, and uses LR (a typical model used in medicine) to explain the selected important features, providing an important preliminary basis for further in-depth research of CSVD.

Our results show that platelet indices and red blood cell indices are potentially related to the imaging phenotypes of CSVD including WMLs, lacunes, microbleeds, atrophy, and EPVS. The high PDW indicates that the destruction of platelets may exceed the normal range, leading to peripheral blood thrombocytopenia, thrombotic diseases, etc. The low P-LCR refers to the low proportion of large platelets in the total number of platelets. The common causes are thrombocytosis and macrothrombocytopenia. The most common pathological factor of high platelets is

Table 6 Regression results for microbleeds.

Marker	OR	95% CI	
		Lower limit	Upper limit
Hematocrit***	0.9782	0.9740	0.9824
PDW**	1.0416	1.0152	1.0687
Urobilinogen	—	—	—
+-***	0.3904	0.3359	0.4534
1+	0.6275	0.3384	1.1754
2+	0.7573	0.2477	2.5698
3+	0.3552	0.0581	2.1502
Indirect bilirubin***	1.1078	1.0573	1.1609
Total bilirubin	0.9762	0.9454	1.0081
Specific gravity of urine***	0.0000	0.0000	0.0000
Glucose***	0.9302	0.9073	0.9535
Eosinophils/100 leukocytes***	1.1192	1.0872	1.1524
Potassium***	0.7789	0.6898	0.8790
ALP***	1.0050	1.0027	1.0074

Note: ***P<0.001 and **P<0.01.

Table 7 Regression results for atrophy.

Marker	OR	95% CI	
		Lower limit	Upper limit
RDW***	1.0751	1.0686	1.0817
Creatinine***	1.0066	1.0052	1.0080
Cholinesterase***	0.9999	0.9999	0.9999
ALT***	0.9934	0.9923	0.9944
Platelets***	0.9986	0.9983	0.9990
WBC	0.9793	0.9567	1.0025
Urate***	1.0007	1.0004	1.0009
BUN***	1.0482	1.0357	1.0609
Neutrophils***	0.9518	0.9295	0.9745
Hematocrit***	1.0093	1.0081	1.0106

Note: ***P<0.001.

Table 8 Regression results for EPVS.

Marker	OR	95% CI	
		Lower limit	Upper limit
Hematocrit***	0.9716	0.9597	0.9832
RDW**	0.9304	0.8864	0.9755
Phosphorus**	0.2674	0.1111	0.6294
Total bilirubin***	1.0573	1.0240	1.0929
LDH*	0.9967	0.9935	0.9998

Note: ***P<0.001, **P<0.01, and *P<0.05.

infection, and more platelets will increase the risk of vascular embolism. The high PMV indicates that the patient may have a disease of the blood system[40−42]. Hematocrit is an important indicator to reflect the state of red blood cells. Low hematocrit indicates possible anemia. The low RDW indicates that the volume and size of red blood cells are relatively uniform. Red blood cells related indicators need to be combined with other indicators for comprehensive clinical judgment[43].

There are other meaningful markers. For example, Ryu et al[6]. reported higher levels of ALP are independently associated with WMH and cerebral infarct, but not with cerebral microbleeds. Liu et al[44]. indicated that high ALP levels in relation to microbleeds in acute ischemic stroke patients. The reasons for the differences between our results and those of these studies mainly include two aspects: different study patients and different analysis methods. In this study, ALP does not enter the top 10 important

markers for WMLs and lacunes. Therefore, if ALP is included in LR of WMLs and lacunes, it may be statistically significant.

To the best of our knowledge, we are the first to apply machine learning models based on OMOP-CDM to study imaging subgroups of CSVD. RF is a classifier that uses multiple trees to train and predict samples, which can process data with very high dimensions (many features), and does not need to reduce dimensions. It can judge the importance of features and the interaction between different features, but it is not easy to explain. The accuracy of RF in Karel et al[24].'s study ranges from 0.6520 to 0.7830, which is comparable to the results of this study. But their research outcomes and included features are completely different from those of this study. The most important thing is that RF performs the best in classification compared with SVM and XGBoost. This is the reason for choosing RF in this study. LR is simple, easy to understand, and very interpretable. From the weight of features, we can see the impact of different features on the final results. However, we cannot use LR to solve nonlinear problems. LR itself cannot select features. Based on the low cost data, this study uses machine leaning models to select important markers for phenotypic subgroups of CSVD imaging.

This study also has limitations. It is a retrospective observational study, which is likely to be affected by unmeasured and unnoticed bias and confounding factors. However, some hypotheses can be quickly obtained from this study, which provides an important preliminary basis for clinical randomized controlled trials.

5 Conclusion

This study proposes a new analytical framework to select important clinical markers for CSVD with machine learning based on a common data model, which has low cost, fast speed, large sample size, and continuous data sources. First, we use PSM to obtain a balanced dataset, use RF to classify imaging subgroups, and select features. Then, we input the top 10 important markers into stepwise LR to obtain OR and 95% CI. Our results find that there are differences and similarities in the important markers of each imaging subgroups of CSVD, with hematocrit being the top 10 markers for all 5 imaging subgroups. There is a need for multi-center data to continue exploring and verifying the effectiveness of selected clinical markers in guiding clinical practice.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 72204169 and 81825007), Beijing Outstanding Young Scientist Program (No. BJJWZYJH01201910025030), Capital's Funds for Health Improvement and Research (No. 2022-2-2045), National Key R&D Program of China (Nos. 2022YFF1501500 2022YFF1501501, 2022YFF1501502, 2022YFF1501503, 2022YFF1501504, and 2022YFF1501505), Youth Beijing Scholar Program (No. 010), Beijing Laboratory of Oral Health (No. PXM2021_014226_000041), Beijing Talent Project-Class A: Innovation and Development (No. 2018A12), National Ten-Thousand Talent Plan-Leadership of Scientific and Technological Innovation, and National Key R&D Program of China (Nos. 2017YFC1307900 and 2017YFC1307905).

References

[1] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, et al., Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, *Lancet Neurol.*, vol. 12, no. 8, pp. 822–838, 2013.

[2] D. Liu, X. Cai, Y. Yang, S. Wang, D. Yao, L. Mei, J. Jing, S. Li, H. Yan, X. Meng, et al., Associations of life's simple 7 with cerebral small vessel disease, *Stroke*, vol. 53, no. 9, pp. 2859–2867, 2022.

[3] H. Chen, Y. Pan, L. Zong, J. Jing, X. Meng, Y. Xu, H. Yan, X. Zhao, L. Liu, H. Li, et al., Cerebral small vessel disease or intracranial large vessel atherosclerosis may carry different risk for future strokes, *Stroke Vasc. Neurol*., vol. 5, no. 2, pp. 128–137, 2020.

[4] L. Jiang, X. Cai, D. Yao, J. Jing, L. Mei, Y. Yang, S. Li, A. Jin, X. Meng, H. Li, et al., Association of inflammatory markers with cerebral small vessel disease in community-based population, *J. Neuroinflammation*, vol. 19, no. 1, p. 106, 2022.

[5] Y. Gao, D. Li, J. Lin, A. M. Thomas, J. Miao, D. Chen, S. Li, and C. Chu, Cerebral small vessel disease: Pathological mechanisms and potential therapeutic targets, *Front. Aging Neurosci*., vol. 14, p. 961661, 2022.

[6] W. S. Ryu, S. H. Lee, C. K. Kim, B. J. Kim, H. M. Kwon, and B. W. Yoon, High serum alkaline phosphatase in relation to cerebral small vessel disease, *Atherosclerosis*, vol. 232, no. 2, pp. 313–318, 2014.

[7] H. B. Lee, J. Kim, S. H. Kim, S. Kim, O. J. Kim, and S. H. Oh, Association between serum alkaline phosphatase level and cerebral small vessel disease, *PLoS One*, vol. 10, no. 11, p. e0143355, 2015.

[8] X. Piao, Z. Jie, and W. Yue, Serum alkaline phosphatase level is correlated with the incidence of cerebral small

vessel disease, *Clin. Invest. Med.*, vol. 42, no. 1, pp. E47–E52, 2019.

[9] M. Wada, H. Nagasawa, K. Kurita, S. Koyama, S. Arawaka, T. Kawanami, K. Tajima, M. Daimon, and T. Kato, Cerebral small vessel disease and C-reactive protein: Results of a cross-sectional study in community-based Japanese elderly, *J. Neurol. Sci.*, vol. 264, nos. 1 & 2, pp. 43–49, 2008.

[10] S. Mitaki, A. Nagai, H. Oguro, and S. Yamaguchi, C-reactive protein levels are associated with cerebral small vessel-related lesions, *Acta Neurol. Scand.*, vol. 133, no. 1, pp. 68–74, 2016.

[11] A. Hassan, B. J. Hunt, M. O'Sullivan, R. Bell, R. D'Souza, S. Jeffery, J. M. Bamford, and H. S. Markus, Homocysteine is a risk factor for cerebral small vessel disease, acting via endothelial dysfunction, *Brain*, vol. 127, no. 1, pp. 212–219, 2004.

[12] K. W. Nam, H. M. Kwon, H. Y. Jeong, J. H. Park, H. Kwon, and S. M. Jeong, Serum homocysteine level is related to cerebral small vessel disease in a healthy population, *Neurology*, vol. 92, no. 4, pp. e317–e325, 2019.

[13] Y. Cao, N. Su, D. Zhang, L. Zhou, M. Yao, S. Zhang, L. Cui, Y. Zhu, and J. Ni, Correlation between total homocysteine and cerebral small vessel disease: A Mendelian randomization study, *Eur. J. Neurol.*, vol. 28, no. 6, pp. 1931–1938, 2021.

[14] M. Wada, H. Nagasawa, K. Kurita, S. Koyama, S. Arawaka, T. Kawanami, K. Tajima, M. Daimon, and T. Kato, Microalbuminuria is a risk factor for cerebral small vessel disease in community-based elderly subjects, *J. Neurol. Sci.*, vol. 255, nos. 1&2, pp. 27–34, 2007.

[15] P. W. Chung, K. Y. Park, J. M. Kim, D. W. Shin, M. S. Park, Y. J. Chung, S. Y. Ha, S. W. Ahn, H. W. Shin, Y. B. Kim, et al., 25-hydroxyvitamin D status is associated with chronic cerebral small vessel disease, *Stroke*, vol. 46, no. 1, pp. 248–251, 2015.

[16] S. E. Park, H. Kim, J. Lee, N. K. Lee, J. W. Hwang, J. J. Yang, B. S. Ye, H. Cho, H. J. Kim, Y. J. Kim, et al., Decreased hemoglobin levels, cerebral small-vessel disease, and cortical atrophy: Among cognitively normal elderly women and men, *Int. Psychogeriatr.*, vol. 28, no. 1, pp. 147–156, 2016.

[17] A. Vilar-Bergua, I. Riba-Llena, N. Ramos, X. Mundet, E. Espinel, A. López-Rueda, E. Ostos, D. Seron, J. Montaner, and P. Delgado, Microalbuminuria and the combination of MRI markers of cerebral small vessel disease, *Cerebrovasc. Dis.*, vol. 42, nos. 1&2, pp. 66–72, 2016.

[18] J. Kim, S. J. Yoon, M. H. Woo, S. H. Kim, N. K. Kim, J. Kim, O. J. Kim, and S. H. Oh, Differential impact of serum total bilirubin level on cerebral atherosclerosis and cerebral small vessel disease, *PLoS One*, vol. 12, no. 3, p. e0173736, 2017.

[19] Z. G. Yin, Q. S. Wang, K. Yu, W. W. Wang, H. Lin, and Z. H. Yang, Sex differences in associations between blood lipids and cerebral small vessel disease, *Nutr. Metab. Cardiovasc. Dis.*, vol. 28, no. 1, pp. 28–34, 2018.

[20] K. W. Nam, H. M. Kwon, H. Y. Jeong, J. H. Park, H. Kwon, and S. M. Jeong, High triglyceride-glucose index is associated with subclinical cerebral small vessel disease in a healthy population: A cross-sectional study, *Cardiovasc. Diabetol.*, vol. 19, no. 1, p. 53, 2020.

[21] J. Kang, W. Luo, C. Zhang, Y. Ren, L. Cao, J. Wu, and H. Li, Positive association between serum insulin-like growth factor-1 and cognition in patients with cerebral small vessel disease, *J. Stroke Cerebrovasc. Dis.*, vol. 30, no. 7, p. 105790, 2021.

[22] M. Chu, Y. Cai, J. Zhong, Y. Qian, Y. Cen, M. Dou, G. Chen, B. Sun, and X. Lu, Subclinical hypothyroidism is associated with basal Ganglia enlarged perivascular spaces and overall cerebral small vessel disease load, *Quant. Imaging Med. Surg.*, vol. 12, no. 2, pp. 1475–1483, 2022.

[23] T. Oberheiden, C. Blahak, X. D. Nguyen, M. Fatar, E. Elmas, N. Morper, C. E. Dempfle, H. Bäzner, M. Hennerici, M. Borggrefe, et al., Activation of platelets and cellular coagulation in cerebral small-vessel disease, *Blood Coagul. Fibrinolysis*, vol. 21, no. 8, pp. 729–735, 2010.

[24] M. F. A. Karel, M. G. C. H. Roosen, B. M. E. Tullemans, C. E. Zhang, J. Staals, J. M. E. M. Cosemans, and R. R. Koenen, Characterization of cerebral small vessel disease by neutrophil and platelet activation markers using artificial intelligence, *J. Neuroimmunol.*, vol. 367, p. 577863, 2022.

[25] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, and M. N. Zozus, Evaluating common data models for use with a longitudinal community registry, *J. Biomed. Inform.*, vol. 64, pp. 333–341, 2016.

[26] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, A survey on missing data in machine learning, *J. Big Data*, vol. 8, no. 1, p. 140, 2021.

[27] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni, Missing value imputation approach for mass spectrometry-based metabolomics data, *Sci. Rep.*, vol. 8, no. 1, p. 663, 2018.

[28] A. Bihorac, T. Ozrazgat-Baslanti, A. Ebadi, A. Motaei, M. Madkour, P. M. Pardalos, G. Lipori, W. R. Hogan, P. A. Efron, F. Moore, et al., MySurgeryRisk: Development and validation of a machine-learning risk algorithm for major complications and death after surgery, *Ann. Surg.*, vol. 269, no. 4, pp. 652–662, 2019.

[29] K. Morita, Introduction to multiple imputation, *Ann. Clin. Epidemiol.*, vol. 3, no. 1, pp. 1–4, 2021.

[30] M. Jakubowski, Latent variables and propensity score matching: A simulation study with application to data from the Programme for International Student Assessment in Poland, *Empir. Econ.*, vol. 48, no. 3, pp. 1287–1325, 2015.

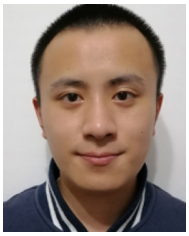
[31] L. Lan, Q. Guo, Z. Zhang, W. Zhao, X. Yang, H. Lu, Z. Zhou, and X. Zhou, Classification of infected necrotizing pancreatitis for surgery within or beyond 4 weeks using machine learning, *Front. Bioeng. Biotechnol.*, vol. 8, p. 541, 2020.

[32] N. Shi, L. Lan, J. Luo, P. Zhu, T. R. W. Ward, P. Szatmary, R. Sutton, W. Huang, J. A. Windsor, X. Zhou, et al., Predicting the need for therapeutic intervention and mortality in acute pancreatitis: A two-center international

study using machine learning, *J. Pers. Med.*, vol. 12, no. 4, p. 616, 2022.
[33] L. Breiman, Random forests, *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[34] Y. Feng, Support vector machine for stroke risk prediction, *Highlights Sci. Eng. Technol.*, vol. 38, pp. 917–923, 2023.
[35] S. Narayanpethkar, M. Rishitha, S. Chandana, and D. T. V. Saradhi, Detection of Parkinson's disease using XGBOOST algorithm, *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 12, pp. 1576–1590, 2022.
[36] R. Díaz-Uriarte and S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, vol. 7, p. 3, 2006.
[37] S. Bernard, L. Heutte, and S. Adam, Influence of hyperparameters on random forest accuracy, in *Proc. Int. Workshop on Multiple Classifier Systems*, Reykjavik, Iceland, 2009, pp. 171–180.
[38] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs, Random forests for genetic association studies, *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, p. 32, 2011.
[39] A. Worster, J. Fan, and A. Ismaila, Understanding linear and logistic regression analyses, *CJEM*, vol. 9, no. 2, pp. 111–113, 2007.
[40] V. Wiwanitkit, Plateletcrit, mean platelet volume, platelet distribution width: Its expected values and correlation with parallel red blood cell parameters, *Clin. Appl. Thromb. Hemost.*, vol. 10, no. 2, pp. 175–178, 2004.
[41] Y. U. Budak, M. Polat, and K. Huysal, The use of platelet indices, plateletcrit, mean platelet volume and platelet distribution width in emergency non-traumatic abdominal surgery: A systematic review, *Biochem. Med.*, vol. 26, no. 2, pp. 178–193, 2016.
[42] K. Pogorzelska, A. Krętowska, M. Krawczuk-Rybak, and M. Sawicka-Żukowska, Characteristics of platelet indices and their prognostic significance in selected medical condition-a systematic review, *Adv. Med. Sci.*, vol. 65, no. 2, pp. 310–315, 2020.
[43] P. R. Sarma, Red cell indices, in *Clinical Methods*: *The History, Physical, and Laboratory Examinations. 3rd ed.* H. K. Walker, W. D. Hall, and J. W. Hurst, eds. Bethesda, MD, USA: National Library of Medicine, 1990.
[44] J. Liu, D. Wang, J. Li, Y. Xiong, B. Liu, C. Wei, S. Wu, and M. Liu, High serum alkaline phosphatase levels in relation to multi-cerebral microbleeds in acute ischemic stroke patients with atrial fibrillation and/or rheumatic heart disease, *Curr. Neurovasc. Res.*, vol. 13, no. 4, pp. 303–308, 2016.

Lan Lan received the PhD degree in epidemiology and health statistics from Sichuan University in 2018. She is currently a postdoctoral researcher with IT Center, Beijing Tiantan Hospital, Capital Medical University. Her research interests include healthcare big data, medical informatics, and medical artificial intelligence.

Tingting Wang received the BM degree from Xuzhou Medical University in 2014, the MM and PhD degrees from Capital Medical University in 2017 and 2020, respectively. She is currently a postdoctoral researcher at Department of Neurology, Beijing Tiantan Hospital, Capital Medical University. Her research interests include medical imaging analysis of cerebral small vessel disease.



Jiawei Luo received the MS degree in public health from West China School of Public Health, Sichuan University in 2015. He is currently pursuing the PhD degree in medical informatics with West China Hospital/West China School of Medicine, Sichuan University. His research interests include machine learning, neural networks, deep learning and other technologies to optimize clinical pathways, and treatment options.

Rui Li received the PhD degree in health management from Huazhong University of Science and Technology in 2017. He is the director of IT Center, Beijing Tiantan Hospital, Capital Medical University. His current research interests include healthcare big data and hospital information management.

Lingling Jiang received the PhD degree from Beijing Normal University in 2017. She was a postdoctoral researcher with Capital Medical University. She joined China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, in 2022. Her research interests include genetics, biomarkers, pathogenesis, and therapeutic targets of cerebral small vessel disease.

Guoliang Hu received the MD degree from Beijing Institute of Heart, Lung, and Blood Vessel Diseases in 2017. He is currently pursuing the PhD degree in Beijing Tiantan Hospital, Capital Medical University. His research interests include cerebral small vessel diseases, autonomic function, cognitive performance, and heart and brain co-morbidity.

Zhiwei Ji received the PhD degree from Tongji University in 2016. He is currently a professor with College of Artificial Intelligence, Nanjing Agricultural University (NJAU), China. He is currently the director of Center for Data Science and Intelligent Computing at NJAU. Prior to this position, he was an assistant professor at University of Texas Health Science Center at Houston (UTHealth), USA. He has been working on systems biology, bioinformatics, pattern recognition, and big data analysis and modeling for more than ten years. He has authored or co-authored more than 50 referred papers published in the world-renowned academic journals, such as *IEEE Transactions on Systems*, *Man*, *and Cybernetics*: *Systems*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *IEEE Systems Journal*, *Information Sciences*, and *PLoS Computational Biology*.

Yilong Wang received the PhD degree from Capital Medical University in 2007. He is the chief physician and professor with Beijing Tiantan Hospital, Capital Medical University, a recipient of the National Science Fund for Distinguished Young Scholars, and the "Ten Thousand People" Program of Organization Department of Central Committee of the CPC. His main research direction is clinical and basic research of cerebral small vessel disease. As the first/corresponding author, he has published more than 100 original papers in *JAMA*, *BMJ*, *Lancet Neurology*, *Circulation*, and other journals