URSI AP-RASC, Sydney, Australia, 17 – 22 August 2025

Machine Learning-Driven Prediction of Cerebral Small Vessel Disease Based on Electromagnetic Device Data

Benben Wang[(1), (2)], Jianwei Xu[(3), *],Yuefei Yan[(1), (2) , *], Huizhen Lu[(4)] , Ke Zang[(1)] , Chengliang Zhang[(4)], Hao Wang[(1), (2)], Chuanliu Wang[(4) *], Congsi Wang[(1), (2)]

(1)State Key Laboratory of Electromechanical Integrated Manufacturing of High-performance Electronic Equipments, Xidian University, Xi'an, China

(2)Guangzhou Institute of Technology, Xidian University, Guangzhou, China

(3)Quzhou Hospital of Traditional Chinese Medicine, Quzhou, China

(4)The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou, China

*Corresponding author

Abstract

With the continuous development of modern medicine, electromagnetic devices have been widely used in the clinical diagnosis of various diseases, which has had a profound impact on human health and quality of life. However, current predictive methods for cerebral small vessel disease based on electromagnetic device measurement data are still limited and mostly constrained by traditional statistical methods, making it difficult to fully exploit complex data patterns. To address this issue, this study proposed a prediction method for cerebral small vessel disease based on ensemble learning. Experimental result showed that the proposed model predicts an AUC value of up to 0.919, demonstrating the enormous potential of machine learning and ensemble learning methods in predicting cerebral small vessel disease. This research provides a new technological path for the application of electromagnetic device data in the early detection and diagnosis of cerebral small vessel disease.

1. Introduction

Cerebral small vessel disease (CSVD) refers to a series of clinical, imaging, and pathological syndromes caused by damage to cerebral small arteries, small perforating arteries, capillaries, and small veins. CSVD generally characterized on cranial magnetic resonance imaging (MRI) by features such as lacunar infarctions (LI), white matter hyperintensities (WMH), enlarged perivascular spaces (EPVS), and cerebral microbleeds (CMB) [1]. Clinical manifestations include stroke, cognitive impairment, emotional disorders, gait abnormalities, and urinary abnormalities. CSVD is a major disease with systemic and cerebral impacts, with a wide range of affected populations worldwide. Most patients have chronic onset and progressive exacerbation, such as cognitive decline, mental abnormalities, gait instability, etc., which seriously affect their quality of life [2].

In recent years, with the continuous advancement of medical technology, electromagnetic equipment has been widely used in the detection and diagnosis of various diseases. For example, electromagnetic wave signals have been used to detect cerebrovascular disease, cardiovascular disease, Parkinson's disease, Alzheimer's disease, and so on. At the same time, machine learning applications have experienced exponential growth and accelerated innovation in recent years [3]. Compared with traditional statistical methods, ML algorithms have fewer constraints on data and can effectively model complex datasets [4], thus their applications in the medical field are gradually increasing. In terms of predicting cerebral small vessel disease based on electronic cases, Elisa analyzed subtypes of cerebral small vessel disease and focused on the progress of machine learning strategies in cerebral small vessel disease [5]. Le Zhang establish a risk prediction model for clinical diagnosis and treatment of cerebral small vessel disease. The logistic multivariable regression model was used to screen out the independent risk factors of cognitive dysfunction in patients with CSVD, and the nomogram of cognitive dysfunction in patients with CSVD was constructed based on the results of the logistic multivariable regression analysis [6]. Fangfang Zhu establish and evaluate a prognostic model for the severity of cerebral small vessel disease based on the nomogram [7]. However, There is a lack of research on the applicability of ensemble machine learning methods for CSVD prediction, particularly in the context of small-sample, high-dimensional data.

In the field of CSVD comorbidities, disease models based on risk factors have also made notable progress, providing valuable references for CSVD prediction. Xuemeng Li utilized data from the 2017 national stroke screening to develop stroke risk classification models based on machine learning algorithms, thereby improving classification efficiency [8]. Elias Dritsas designed a robust framework for the long-term risk prediction of stroke occurrence [9]. Ali F

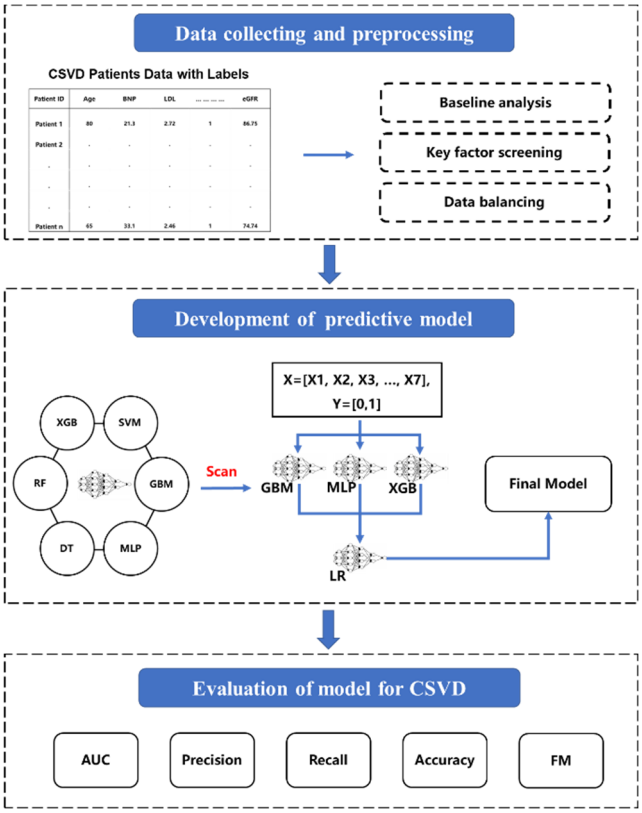

Figure 1. The proposed prediction process for CSVD

proposed a heart disease prediction method that integrates wearable device data and electronic medical record data. The process involved constructing multiple models, including a data collection layer, data fusion and extraction layer, data preprocessing layer, ensemble learning layer, and recommendation layer, achieving a prediction accuracy of 98.5% [10].

In summary, with the development of neuroimaging, scholars have significantly improved their understanding of the diagnosis and treatment of cerebral small vessel disease. However, current research on CSVD prediction based on electromagnetic equipment data is still relatively limited, mostly relying on traditional statistical methods. These methods have limitations in processing high-dimensional and nonlinear data, making it difficult to fully explore potential feature patterns in the data. This study aims to address the significant demand for diagnosis and treatment of cerebral small vessel disease. A multidimensional feature screening method was proposed to identify early-intervention risk factors for cerebral small vessel disease; Suitable machine learning models were screened for predicting cerebral small vessel disease. Based on ensemble learning methods a cerebral small vessel disease prediction model was constructed; The proposed model was validated using multiple evaluation metrics, providing a solid theoretical foundation and technical support for early warning and health management of cerebral small vessel disease. The detailed process has been illustrated in Figure 1.

2. Intelligent prediction method for CSVD

2.1 Data collection

A total of 417 patients diagnosed with CSVD and 189 patients with other neurological disorders were hospitalized in Department of Neurology at Quzhou Hospital, Affiliated to Wenzhou Medical University between February 2022 and June 2024 were selected. The collected CSVD data included basic clinical inquiry data, laboratory test data, and imaging examination data, all of which have been screened according to standardized inclusion and exclusion criteria. Imaging examinations were presented in the form of reports as the criteria for doctors to determine CSVD. We collected data on 27 risk factors, including age, gender, years of schooling, BMI, hypertension, diabetes, LDL, BNP, heart disease, merge previous stroke, merge new stroke, and homocysteine, among others. A baseline analysis was conducted to assess the statistical significance of the data.

2.2 Key risk factor screening

Before conducting the screening of key risk factors, we first preprocessed the data, including filling in missing values and normalizing the data. For missing values, we used the mean imputation method to reduce data bias and maintain data integrity. We employed the Min-Max Normalization technique to improve the stability of model training and the accuracy of analysis. These data preprocessing steps help improve the generalization ability of the model and the credibility of the results.

This study proposed a multidimensional feature screening method to sort features and identify an optimal set of risk factors. The multidimensional feature screening method proposed in this study integrates the strengths of Random Forest, Lasso Regression, Logistic Regression, and Boruta to identify key features in small-sample, high-dimensional datasets. The methods were synthesized using the following scoring formula:

$$Combined\_Score_i = \frac{S_{1i}^{Norm} + S_{2i}^{Norm} + S_{3i}^{Norm} + S_{4i}^{Norm} * 0.5}{\sum_{i=1}^{26}(S_{1i}^{Norm} + S_{2i}^{Norm} + S_{3i}^{Norm} + S_{4i}^{Norm} * 0.5)} \quad (1).$$

Where $Combined\_Score_i$ represents the importance score of the i-th feature; S_{1i}^{Norm}, S_{2i}^{Norm}, S_{3i}^{Norm}, S_{4i}^{Norm} represent the normalized importance scores of the i-th feature obtained from Random Forest, Lasso Regression, Logistic Regression, and Boruta, respectively, with i ranging from 1 to 25. If $Combined\_Score_i \geq 0.04$, it is considered that this factor has an extremely remarkable impact on CSVD.

2.3 Development of predictive model

Before building the model, we considered using the SMOTE method for data balancing, generating new negative samples

Table 1 Performance metrics.

Name	Description
Accuracy(Acc)	$\frac{TP+TN}{FP+TN+FN+TP}$
Precision(Pre)	$\frac{TP}{FP+TP}$
Recall(Rec)	$\frac{TP}{TP+FN}$
$F\text{-}Measure(FM)$	$\frac{2*P*R}{R+P}$

through interpolation instead of simply copying existing negative samples, in order to avoid overfitting of the model.

The model construction was based on the selected features and balanced data of CSVD. In terms of algorithm selection, seven classical machine learning models were chosen, including Logistic Regression(LR), Decision Tree(DT), Random Forest(RF), Support Vector Classifier(SVM), Gradient Boosting(GBM), XGBoost(XGB), and MLP Classifier(MLP), considering the advantages and disadvantages of each model. Predictive models were then constructed to evaluate the performance of different classifiers on the current CSVD dataset. In order to ensure the robustness of the model evaluation, five-fold cross-validation was used for model validation, which helps to maintain consistent distribution of target variables in each fold dataset during training and validation, thereby reducing bias caused by data imbalance. Finally, the AUC values and accuracy of each model were comprehensively evaluated to evaluate different classifiers, providing a basis for subsequent model selection.

To further enhance the overall performance of the classification model, a stacking method was employed based on the selected models to construct an integrated classification model. As a powerful ensemble learning approach, stacking offers three key advantages: superior model performance, strong interpretability, and suitability for complex data. The fundamental idea of stacking is to first train multiple base models independently, then use their outputs as inputs to a meta-learner, which ultimately generates the final prediction results.

2.4 Evaluation of model for CSVD

The proposed diagnostic prediction model is systematically compared with various mainstream classifiers to comprehensively assess its superior performance. Receiver Operating Characteristic (ROC) curve is used to ensure both high True Positive Rate (TPR) and low False Positive Rate (FPR) simultaneously. Additionally, other metrics, such as Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-Measure (FM), are applied to evaluate the model's overall performance and effectiveness, as shown in Table 1.

Table 2 The comparison chart of multiple models

	AUC	Acc	Pre	Rec	FM
LR	0.855	0.799	0.801	0.800	0.800
DT	0.785	0.780	0.780	0.780	0.780
RF	0.880	0.795	0.795	0.795	0.795
SVM	0.838	0.773	0.776	0.773	0.772
GBM	0.895	0.809	0.809	0.809	0.809
XGB	0.893	0.812	0.812	0.812	0.812
MLP	0.894	0.842	0.843	0.842	0.842
Proposed	0.919	0.855	0.856	0.855	0.855

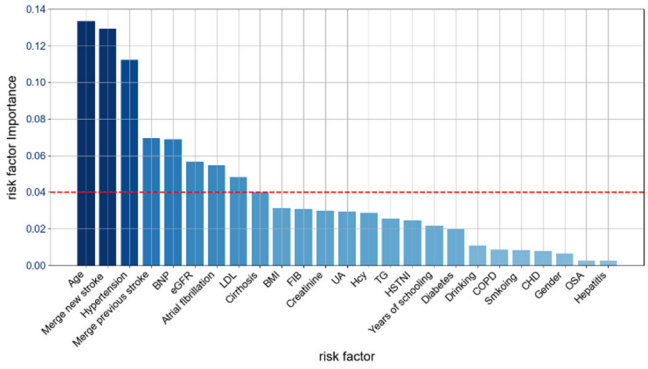

Figure 2. Feature selection and Ranking of CSVD.

3. Results and Discussion

3.1 Key Feature Selection Analysis

Through the proposed method, the significant risk factors of CSVD are identified through screening, and the ranking of risk factors is shown in Figure 2. The ranking of these risk factors is largely consistent with clinical findings, demonstrating the scientific validity of this method. Based on the selection criteria, the key risk factors of CSVD are determined.

$$F_k = \begin{bmatrix} \text{Age, Merge New Stroke, Hypertension, BNP, eGFR,} \\ \text{Merge Previous Stroke, Atrial Fibrillation, LDL} \end{bmatrix}$$

3.2 Development and evaluation of model

Using the above key risk factors of CSVD, we constructed a CSVD diagnostic model based on seven common machine learning models. Among the seven machine learning models, the AUC ranking is MLP > GBM > XGB > RF > LR > SVM > DT, with the top four models significantly outperforming the latter ones. In terms of Accuracy, the ranking is MLP > XGB > GBM > LR > RF > DT > SVM, again with the top four models showing a clear performance advantage.

Considering both factors comprehensively, MLP, GBM, and XGB were selected for performance enhancement based on Stacking method. The proposed model was evaluated using AUC, Precision, Recall, F1-score, and Accuracy. The model achieved an AUC of 0.919, demonstrating excellent

performance in the field of CSVD diagnosis. A detailed comparison of model performance is presented in Table 2.

The selection of these three models is also based on their unique strengths. The MLP method has a powerful nonlinear modeling capability and can approximate any continuous function, making it effective for disease risk prediction and classification tasks. XGB has a well-designed regularization mechanism that effectively prevents overfitting. It offers extensive hyperparameter tuning options, providing strong model flexibility, good generalization ability, and high interpretability. GBM is well-suited for nonlinear data, capable of learning complex feature relationships. It provides fast training speed, supports categorical feature processing, is highly flexible, and demonstrates strong stability. These advantages collectively contribute to the superior predictive performance of the proposed ensemble model.

4. Conclusion

In this article, we propose a prediction model for cerebral small vessel disease based on ensemble learning method, aiming to significantly improve the accuracy and efficiency of clinical diagnosis of cerebral small vessel disease. We have conducted in-depth research on multiple key issues faced in this field, including the following. Firstly, data inclusion and exclusion criteria were set, and the collected data was preprocessed. Baseline analysis was conducted based on the statistical significance of the data; Secondly, a multidimensional feature evaluation and selection method was proposed to address the redundancy of data features in cases of cerebral small vessel disease, and effective measures were taken to address the issue of data imbalance. Thirdly, based on the characteristics of small sample high-dimensional medical data, a machine learning model suitable for predicting cerebral small vessel disease is screened. An integrated machine learning method is used to construct the prediction model, and cross validation is conducted to evaluate its performance. Our proposed model achieved an AUC value of 0.919 and performed well in other aspects, demonstrating the strong practicality of ensemble learning methods in the field of cerebral small vessel disease.

In future, we will utilize more electromagnetic devices, such as wearable ultrasound imaging devices, for intelligent monitoring of cerebral small blood vessels, enabling early detection and intervention of cerebral small vessel disease.

5. Acknowledgements

This work was supported by National Natural Science Foundation of China under No. 52275268 and U23A6017, The Project about Building up "Scientists + Engineers" of Shaanxi Qinchuangyuan Platform under No. 2022KXJ-030, Xidian University Specially Funded Project for Interdisciplinary Exploration (TZJH2024025).

References

[1] J. M. Wardlaw, C. Smith, and M. Dichgans, “Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging,” *Lancet Neurol,* vol. 12, no. 5, pp. 483-97, May 2013, doi: 10.1016/S1474-4422(13)70060-7.

[2] S. M. Kwon et al., "Impact of brain atrophy on 90-day functional outcome after moderate-volume basal ganglia hemorrhage, ” Sci Rep, vol. 8, no. 1, p. 4819, Mar 19 2018, doi: 10.1038/s41598-018-22916-3.

[3] S. S. Al-Zaiti et al., "A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML)," Eur Heart J Digit Health, vol. 3, no. 2, pp. 125-140, Jun 2022, doi: 10.1093/ehjdh/ztac016. [4]

[4] W. Ren, D. Li, J. Wang, J. Zhang, Z. Fu, and Y. Yao, "Prediction and Evaluation of Machine Learning Algorithm for Prediction of Blood Transfusion during Cesarean Section and Analysis of Risk Factors of Hypothermia during Anesthesia Recovery," Comput Math Methods Med, vol. 2022, p. 8661324, 2022, doi: 10.1155/2022/8661324.

[5] E. Cuadrado-Godia et al., "Cerebral Small Vessel Disease: A Review Focusing on Pathophysiology, Biomarkers, and Machine Learning Strategies," J Stroke, vol. 20, no. 3, pp. 302-320, Sep 2018, doi: 10.5853/jos.2017.02922.

[6] L. Zhang et al., "Analysis of risk factors for the development of cognitive dysfunction in patients with cerebral small vessel disease and the construction of a predictive model," Front Neurol, vol. 13, p. 944205, 2022, doi: 10.3389/fneur.2022.944205.

[7] F. Zhu, J. Yao, M. Feng, and Z. Sun, "Establishment and evaluation of a clinical prediction model for cognitive impairment in patients with cerebral small vessel disease," BMC Neurosci, vol. 25, no. 1, p. 35, Aug 2 2024, doi: 10.1186/s12868-024-00883-y.

[8] X. Li, D. Bian, J. Yu, M. Li, and D. Zhao, "Using machine learning models to improve stroke risk level classification methods of China national stroke screening," BMC Med Inform Decis Mak, vol. 19, no. 1, p. 261, Dec 10 2019, doi: 10.1186/s12911-019-0998-2.

[9] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors (Basel), vol. 22, no. 13, Jun 21 2022, doi: 10.3390/s22134670.

[10] F. Ali et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Information Fusion, vol. 63, pp. 208-222, 2020, doi: 10.1016/j.inffus.2020.06.008.