Heliyon 10 (2024) e24232

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Construction and comparison of short-term prognosis prediction model based on machine learning in acute ischemic stroke

Yinting Xing [a], Yingyu Jin [a], Yanhong Liu [b,*]

[a] Department of Clinical Laboratory, The First Affiliated Hospital of Harbin Medical University, Harbin City, Heilongjiang Province, China
[b] Department of Clinical Laboratory, The Second Affiliated Hospital of Harbin Medical University, Harbin City, Heilongjiang Province, China

ARTICLE INFO

Keywords:
Acute ischemic stroke (AIS)
Machine learning (ML)
Random forest (RF)
Neutrophil multiplied by D-dimer (NDM)

ABSTRACT

Objective: To construct and compared the short-term prognosis prediction models of acute ischemic stroke (AIS) by machine learning (ML).
Methods: Retrospectively study. The group W (mRS≤3) was clustered, and combined with group P (mRS>3) to form the post-clustering dataset for modeling. The "glmnet", "rpart", "xgboost", "randomForest", "neuralnet" packages were used to construct ML models. The accuracy, sensitivity, specificity, positive predict value (PPV), negative predict value (NPV) among the models were compared. Four external clinical datasets were used for external clinical validation. The optimal prediction model was determined by variable screening ability, model visualization, and external clinical validation performance.
Results: The post-clustering dataset contains 139 patients (group W) and 122 patients (group P). The neutrophil multiplied by D-dimer (NDM) has predictive value in all ML prediction models in this study. In the decision tree model, NDM^Q occupies the first tree node, When NDM≤5.62 and the age<74.5, the probability of poor prognosis of AIS is less than 20 %. When NDM>5.62 and accompanied by pneumonia, the incidence of poor prognosis of AIS is about 90 %. In the Random Forest (RF) model, NDM^Q had the highest Gini index. The variable combination screened by the RF model had the best performance in the neural network, and the accuracy, sensitivity, specificity, PPV, and NPV of the external validation were 0.800, 0.774, 0.833, 0.857, and 0.741, respectively. The RF model had the best performance in the external clinical validation datasets, with accuracies of 0.646, 0.697, 0.695, and 0.713, respectively.
Conclusions: NDM shows predictive value for AIS short-term prognosis in all ML models in this study. The optimal model in screening characteristic variables and the performance of in external clinical datasets was RF model. In the analysis of medical data with small sample size and outcome as categorical variables, RF could be used as the main algorithm to build a model.

1. Introduction Background

Acute ischemic stroke (AIS) is a disease with elevated risk and proportion of death, which is associated with large disability-adjusted life year and poor short-term prognosis after onset [1,2]. Global population growth and aging will cause the AIS burden remaining high, increase and become younger [3]. For this, the establishment of a rapid prognostic prediction model could provide early warning to clinics, and the incidence of poor prognosis could be reduced through early detection, diagnosis and intervention.

* Corresponding author.
E-mail address: liuusa2016@163.com (Y. Liu).

https://doi.org/10.1016/j.heliyon.2024.e24232
Received 18 April 2023; Received in revised form 25 November 2023; Accepted 4 January 2024
Available online 6 January 2024

Routine laboratory tests such as neutrophil [4], lymphocyte [5], folic acid [6], glucose [7] may reflect the underlying problems leading to poor prognosis in AIS patients. Clinical information about AIS patients can affect their short-term prognosis. Clinical information, including demographic information, past history, personal history, medication history and onset clinical symptoms, can be easily queried or obtained in patients' medical records. Comprehensive analysis of laboratory project results and clinical information will help improve the performance of the prediction model. At present, the construction methods of machine learning (ML) include LASSO cross-validation regression, decision tree-classification tree, Xgboost, Random Forest (RF) and artificial neural network (ANN), etc. [8] Few studies have combined blood cell analysis, coagulation, lipid profile, immune function, and clinical information to construct rapid prognostic models by ML [9]. Besides, there are limitations in using a single ML method [10]. For the prediction classification problems, it is currently not possible to provide a standard solution for a particular problem, because spurious results need to be avoided. Instead, it is prudent to construct the best model through empirical and trial methods.

In summary, this study retrospectively analyzed the first routine laboratory test items, the first page of course information during hospitalization, and the 3-month follow-up records of AIS patients after admission, and constructed a short-term rapid prognosis model for AIS based on the ML. The optimal model was found by comparing models constructed by different ML methods.

2. Materials and methods

2.1. Enrollment inclusion and exclusion criteria

Patients with ischemic encephalopathy in the First Affiliated Hospital of Harbin Medical University (HMU) from July 2019 to July 2021 were selected for retrospective study. The diagnosis of AIS was based on the Chinese Guidelines for the Diagnosis and Treatment of Acute Ischemic Stroke 2018. Short-term prognosis: all patients were followed up for 3 months to track modified RANKIN Scale (mRS) records, mRS>3 was classified as poor prognosis, because patients with mRS>3 would show severe disability or even death, which would seriously affect the quality of life, combined with previous literature studies, 3 was used as the cut-off value for analysis and modeling in this study.

Inclusion criteria: patients diagnosed with AIS for the first time; age>18 years old; patient or family member sign the subject's informed consent.

Exclusion criteria: those who were age<18 years old; complications include acute myocardial infarction, pulmonary embolism, venous thrombosis, surgery, tumors, diffuse intravascular coagulation, severe infection, tissue necrosis, and other clinical diseases; incomplete clinical data or laboratory project data; withdraw; refuse to be included.

2.2. Collection of clinical information

Through the hospital's electronic medical record system at the First Affiliated Hospital of HMU, comprehensive data including basic information (gender and age), personal history (smoking history, smoking index, and drinking history), past medical history, previous medication uses records, clinical manifestations, and mRS recorded during a 3-month follow-up period were meticulously collected by experienced neurologists. The smoking index (SI) was calculated as the number of cigarettes smoked per day multiplied by the number of years smoked. Given that some AIS cases exhibit a low incidence of prior diseases, complex histories of medication usage, and diverse clinical presentations, those may not be adequately captured in limited medical records quantity-wise; these factors were excluded. Consequently, this study focused solely on analyzing common past medical history elements, previous medication uses records, and clinical manifestations as predictive factors.

The comprehensive medical history encompasses the following conditions: transient ischemic attack (TIA), diabetes, hypertension, coronary atherosclerotic heart disease (hereinafter referred to as heart disease or cardiac disease), atrial fibrillation (AF), pneumonia, hyperlipidemia (HLP), and hyperhomocysteinemia (HHCY). The previous medication usage includes antidiabetic drugs, antihypertensive agents, lipid-lowering drugs (LLDs), and AF treatment (AFT). AF can be categorized into stable-heart rate medications and anticoagulant therapy. Clinical manifestations comprise aphasia and Babinski sign (+). During the follow-up period, all patients received prompt pharmacological intervention within 3-months after discharge.

Meanwhile, this study aimed to develop a 3-month prognosis model for AIS patients, the factors related to post-discharge medication treatment were not considered. According to whether patients with AIS had poor prognosis, they were divided into group P (poor prognosis, mRS>3) and group W (well prognosis, mRS≤3).

2.3. Testing and data collection of routine laboratory items

Routine laboratory testing data for AIS patients tested and collected within 2 h of admission are as follows:

Neutrophil, lymphocyte, monocyte, eosinophil, basophil, hemoglobin, red blood cell distribution width, platelet, platelet distribution width, prothrombin time, prothrombin activity, international normalized ratio, activated partial thrombin time, fibrinogen, D-dimer, homocysteine, albumin, prealbumin, glucose, cholesterol, triglyceride, high density lipoprotein cholesterol, low density lipoprotein cholesterol, ratio of LDL to HDL, apolipoprotein A, apolipoprotein B, ratio of APOA to APOB, lipoprotein a. Based on our previous studies, the neutrophil, lymphocyte and D-dimer were calculated to obtain new variables [11], NLR is ratio of neutrophil to lymphocyte, NDM is neutrophil multiplied by D-dimer. See Supplementary Table 1 for abbreviations and units.

Since neutrophil, lymphocyte and D-dimer had formed new variables through calculation, they would no longer be included in the subsequent modeling in order to reduce the collinearity between variables, and only NLR and NDM would be included. At the same

time, abnormal variables such as LLDs, basophil and PTINR were eliminated to avoid overfitting the model. NLR and NDM were visual binning as follows: NLR^{Q1}:≤2.17, NLR^{Q2}: 2.18–3.14, NLR^{Q3}: 3.15–4.48, NLR^{Q4}: 4.49–8.06, NLR^{Q5}: 8.07+; NDM^{Q1}:≤2.62, NDM^{Q2}: 2.63–5.62, NDM^{Q3}: 5.63–12.47, NDM^{Q4}: 12.48–42.43, NDM^{Q5}: 42.44+.

2.4. Description of the basic data

The measurement data of the research are in skewness distribution, thus expressed as M (P25, P75). Count data is expressed as N (%). The statistical analysis software is IBM SPSS statistics 25.0, R 4.2.0 and RStudio (2022.07.2 + 576). The graphics software is Adobe Photoshop CC 2018 and Adobe Illustrator 2023.

2.5. Data clustering based on K-means

Data pre-processing is based on the real-world of medical data samples. There were a total of 1856 patients with AIS, of whom only 122 had a poor prognosis, a ratio of approximately 1:14 between poor and well prognosis. K-means clustering was used to keep the sample size mostly consistent between groups and ensure the objectivity of the data screening. Group W, which occupies a large proportion in the original dataset, was effectively clustered to ensure the highest intra-cluster similarity and the lowest similarity among the clusters. Data from each cluster were selected according to the 8 % and combined with group P to form the post-clustering dataset.

ML model performance test uses part of data from the one database (train set) for model training and cross-validation, and data from another database (test set) for external-validation. In this study, the post-clustering data was randomly split into two sets, with the train set accounting for 70 % and the test set accounting for 30 %.

2.6. The LASSO cross-validation regression modeling

The “glmnet” package was used to construct LASSO cross-validation regression modeling, and internal and external-validation of the model were done. After the confusion matrix was generated, the accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and AUC were calculated for both train and test sets. The nomogram was drawn to realize the visualization of the model.

2.7. The decision tree-classification tree modeling

The "rpart" package was used for decision tree-classification tree modeling. When referring to the minimum value of xerror, the corresponding tree node is selected. Classification tree modeling was divided into three steps. In the first step, only clinical information was included. In the second step, only the laboratory test results were included. In the third step, the characteristics selected in the two steps above were included, to conduct the final decision tree-classification tree modeling and draw the tree model to realize visualization. After the confusion matrix of test set was generated, the accuracy, sensitivity, specificity, PPV and NPV were calculated for test set.

2.8. The xgboost in screen characteristics

The “xgboost” package was only used for screen characteristics.

2.9. The RF modeling

The “randomForest” package was used to carry out RF modeling. After the confusion matrix was generated, the accuracy, sensitivity, specificity, PPV, NPV and AUC were calculated for both train and test sets, and the ROC curve was plotted.

2.10. ANN modeling

The “neuralnet” package was used to carry out ANN modeling. After data standardization, the characteristics screened by LASSO cross-validation regression/decision tree-classification tree/Xgboost/RF were as the predictive factors to construct the ANN model. After the confusion matrix was generated, the accuracy, sensitivity, specificity, PPV and NPV were calculated for both train and test sets. The optimal model for this study was found by comparing the models and given an objective evaluation.

2.11. External clinical validation of predictive models

In order to verify the clinical application ability of the prediction model and the relationship between model performance and sample size, this study conducted external clinical validation of the prediction model. There were 4 external clinical validation datasets with different sample size. They are the information of patients with ischemic encephalopathy in the First Affiliated Hospital of HMU from July 2018 to June 2019; in the First Affiliated Hospital of HMU from August 2021 to July 2022; in the Second Affiliated Hospital of HMU from July 2019 to July 2021; in the Fourth Affiliated Hospital of HMU from July 2019 to July 2021. Subject inclusion,

exclusion criteria, clinical information and laboratory data collection are the same as before.

The accuracy, sensitivity, specificity, PPV and NPV of LASSO cross-validation model, decision tree-classification tree model and RF model in four external clinical data were calculated by R external validation program, and the accuracy of the models was compared. In order to reflect the application value of the prediction model in the real world, the external data is not cluster analyzed.

3. Results

3.1. Data set construction

After removing the missing values, there were 1734 patients in group W and 122 patients in group P. After K-means clustering, it was ideal to cluster the patients in group W into 4 clusters, with 1264, 159, 236, and 75 in each After 8 % patients in each cluster were randomly selected, 139 patients in group W combined with 122 patients in group P formed the post-clustering dataset, for the subsequent construction of the prediction model (Fig. 1a and b).

The train set (70 %) and test set (30 %) were randomly allocated. It is worth noting that the number of patients in the train and test sets fluctuates within a small range each time the R was used for random assignment of the datasets, but this does not affect the modeling results. In this study, one of the distribution results were selected to show the post-clustering dataset. The number of patients in the train set was 184, including 93 in group W and 91 in group P. The test set consisted of 77 patients, including 46 patients in group W and 31 patients in group P (Supplementary Table 2).

3.2. LASSO cross-validation regression modeling results

The minimum λ was calculated as 0.0248. When confirming the suitable model, λ1se was used, and its value was 0.0643 (Fig. 2a/2b). The following characteristics were selected by the LASSO cross-validation regression model: AGE, PNEUMONIA, BARBINSKI, PTA, HCY, PROALB, APOB, NDM^Q. The results show that NDM^Q is significant in the model, and the OR values are all greater than 4 (Table 1). Nomogram was used for visualization (Fig. 2c). The regression model was internally validated, and the C-statistic was 0.910. After correction, the C-statistic was 0.871. The calibration curve has been drawn (Fig. 2d–e), and it can be seen from the figure and results that the internal-validation and external-validation results were good. The AUC of external-validation was 0.825 (95%CI: 0.732–0.918) (Fig. 2f). The confusion matrix in the train and test sets were shown in Table 2. The performance was shown in Table 3 for both the train and test sets. The accuracy, sensitivity, specificity, PPV and NPV were 0.853, 0.846, 0.860, 0.856, and 0.851, in the train set and 0.753, 0.806, 0.717, 0.658, and 0.846 in the test set, respectively.

3.3. Decision tree-classification tree modeling results

The following three steps were adopted in the characteristic selection. Step 1: Only clinical information involved. The tree node was 2, involving three characteristics, including PNEUMONIA and AGE (Fig. 3a). Step 2: only laboratory test data involved. The tree node was 1, including NDM^Q (Fig. 3b). Step 3: Screened characteristics in the two steps above involved. The tree node was 5, involving three characteristics, including NDM^Q, PNEUMONIA and AGE (Fig. 3c). The confusion matrix for the three tree models was shown in Table 2. The result of confusion matrix operation shows that the prediction model constructed with only laboratory information is the most outstanding in terms of sensitivity, with sensitivity of 0.806, accuracy of 0.727, specificity of 0.674, PPV of 0.625 and NPV of 0.838. The prediction models constructed with only clinical information were average in accuracy, sensitivity and specificity. The accuracy, sensitivity, specificity, PPV, PPV and NPV were 0.740, 0.774, 0.717, 0.649 and 0.825 respectively (**Table 3**).

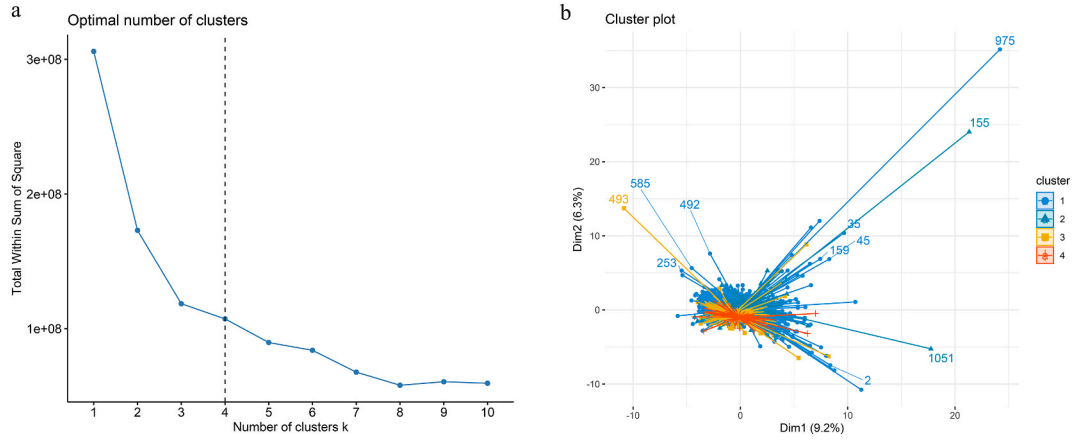

Fig. 1. a: Optimal clustering number diagram, the optimal clustering number is 4. b: Cluster diagram of the 4 clusters for the Group W in the original data.

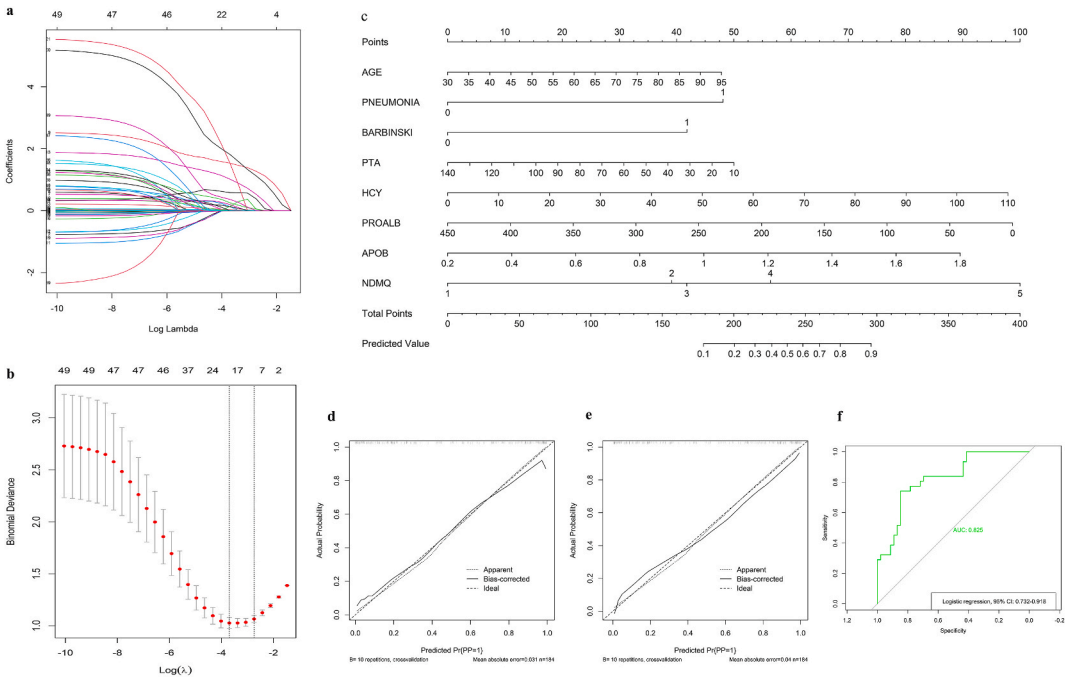

Fig. 2. a: The coefficients of LASSO cross-validation regression. b: LASSO calculated variable to filter λ and calculate the minimum λ and λ1se. c: Nomogram for LASSO cross-validation regression modeling. d: Calibration curve for internal validation of LASSO cross-validation regression model. e: Calibration curve for external validation of LASSO cross-validation regression model. f: ROC curve for external validation of LASSO cross-validation regression model.

Table 1
Regression model constructed by LASSO cross validation in data after clustering.

Characteristic	OR	95 % CI	*P*
AGE	1.03	0.99, 1.07	0.200
PNEUMONIA	6.08	1.94, 22.40	0.003
BARBINSKI	4.80	1.97, 12.50	<0.001
PTA	0.99	0.96, 1.01	0.200
HCY	1.03	1.00, 1.07	0.048
PROALB	0.99	0.98, 1.00	0.044
APOB	8.17	1.60, 47.30	0.014
NDM^Q			
≤ 2.62	–	–	
2.63–5.62	4.34	0.93, 29.30	0.085
5.63–12.47	4.80	0.99, 32.60	0.071
12.48–42.43	8.30	1.64, 58.30	0.017
42.44+	42.60	6.94, 389.00	<0.001

Note: "OR": Odds ratio. "CI": confidence interval.

NDM^Q occupies the first tree node in the classification tree model established by screening feature variables, suggesting that NDM^Q plays an important role in the classification tree prediction model. When NDM<5.62 and the age<74.5, the probability of poor prognosis of AIS is less than 20 %. When NDM>5.62 and accompanied by pneumonia, the incidence of poor prognosis in AIS is very high, about 90 % (Fig. 3c).

3.4. Xgboost screened characteristics

In this study, all characteristics of the clustered data were included, and Xgboost was used to screen characteristics (Fig. 4). The results showed that the following characteristics were prominent in the Xgboost method, namely, NDM^Q, HCY, PNEUMONIA, PROALB, PTA, and NLR^Q. The feature variables screened by the Xgboost are the same as those of the decision tree-classification tree, NDM^Q and PNEUMONIA, suggesting the important influence of these two variables on the prognosis prediction model of AIS. The variables selected by the Xgboost would be validated by the ANN and compared between models.

Table 2
The confusing matrix of ML models.

Model	Set	Prognosis	Well (Predict)	Poor (Predict)
LASSO cross-validation	Train set	Well (True)	80	13
		Poor (True)	14	77
	Test set	Well (True)	33	13
		Poor (True)	6	25
Classification tree				
Clinical information only[1]	Test set	Well (True)	33	13
		Poor (True)	9	22
Laboratory test only[2]	Test set	Well (True)	31	15
		Poor (True)	6	25
Screening characteristics[3]	Test set	Well (True)	33	13
		Poor (True)	7	24
RF	Train set	Well (True)	74	20
		Poor (True)	19	71
	Test set	Well (True)	29	17
		Poor (True)	6	25
ANN				
LASSO cross validation regression	Test set	Well (True)	15	5
		Poor (True)	9	26
Decision tree-classification tree	Test set	Well (True)	28	13
		Poor (True)	5	9
Xgboost	Test set	Well (True)	21	6
		Poor (True)	12	16
Random Forest	Test set	Well (True)	20	4
		Poor (True)	7	24

Note: [1]: Clinical information only included the Characteristics as follows: SEX, AGE, TIA, DIEBETES, HYPERTENSION, CARDIAC, AF, PNEUMONIA, HLP, HHCY, APHASIA, BARBINSKI, SMOKE, SI, DRINKING, ANTIDIABETIC, HEPOTENSOR, AFT. [2]: Laboratory test only included the Characteristics as follows: MONO, ESO, HGB, RDW, PLT, PDW, PT, PTA, APTT, FIB, HCY, ALB, PROALB, GLU, CHOL, TG, HDL, LDL, LHR, APOA, APOB, ABR, LPa, NLR^{Q}, $NDM^{Q.}$ [3]: Screening characteristics included the Characteristics as follows: NDM^{Q}, AGE, PNEUMONIA.

Table 3
The performance of ML models.

Model	Set	Accuracy	Sensitivity	Specificity	PPV	NPV
LASSO cross-validation	Train set	0.853	0.846	0.860	0.856	0.851
	Test set	0.753	0.806	0.717	0.658	0.846
Classification tree						
Clinical information only[1]	Test set	0.714	0.710	0.717	0.629	0.786
Laboratory test only[2]	Test set	0.727	0.806	0.674	0.625	0.838
Screening characteristics[3]	Test set	0.740	0.774	0.717	0.649	0.825
RF	Train set	0.788	0.789	0.787	0.780	0.796
	Test set	0.701	0.806	0.630	0.595	0.829
ANN						
LASSO-ANN[4]	Test set	0.745	0.743	0.750	0.839	0.625
Decision tree - ANN[5]	Test set	0.673	0.643	0.683	0.409	0.848
Xgboost-ANN[6]	Test set	0.673	0.571	0.778	0.727	0.636
RF-ANN[7]	Test set	0.800	0.774	0.833	0.857	0.741

Note: PPV indicates for positive predictive value, NPV indicates for negative predictive value. [1]: Clinical information only included the Characteristics as follows: SEX, AGE, TIA, DIEBETES, HYPERTENSION, CARDIAC, AF, PNEUMONIA, HLP, HHCY, APHASIA, BARBINSKI, SMOKE, SI, DRINKING, ANTIDIABETIC, HEPOTENSOR, AFT. [2]: Laboratory test only included the Characteristics as follows: MONO, ESO, HGB, RDW, PLT, PDW, PT, PTA, APTT, FIB, HCY, ALB, PROALB, GLU, CHOL, TG, HDL, LDL, LHR, APOA, APOB, ABR, LPa, NLR^{Q}, $NDM^{Q.}$ [3]: Screening characteristics included the Characteristics as follows: NDM, AGE, PNEUMONIA. PPV indicates for positive predictive value, NPV indicates for negative predictive value. [4]: ANN model with the characteristics included AGE, PNEUMONIA, BARBINSKI, PTA, HCY, PROALB, APOB, NDM^{Q}. [5]: ANN model with the characteristics included NDM^{Q}, PNEUMONIA and AGE. [6]: ANN model with the characteristics included NDM^{Q}, HCY, PNEUMONIA, PROALB, PTA, and NLR^{Q}. [7]: ANN model with the characteristics included NDM^{Q}, NLR^{Q}, PROALB, MONO, PT, HCY, and PNEUMONIA.

3.5. RF modeling results

When the tree model suitable for operation in the train set was 49, the error was minimal (Fig. 5a). The accuracy and GINI graph obtained after the RF model was set to 49 (Fig. 5b). The GINI graph showed that NDM^{Q}, NLR^{Q}, PROALB, MONO, PT, HCY and PNEUMONIA were relatively important in the model. The confusion matrix results in the train and test sets were shown in Table 2. The accuracy, sensitivity, specificity, PPV, NPV, and AUC were 0.788, 0.789, 0.787, 0.780, 0.796, and 0.810 in the train set (Fig. 5c), meanwhile, those 0.701, 0.806, 0.630, 0.595, 0.829, and 0.718 respectively (Fig. 5d) (Table 3), indicating that the results of internal and external validation were acceptable.

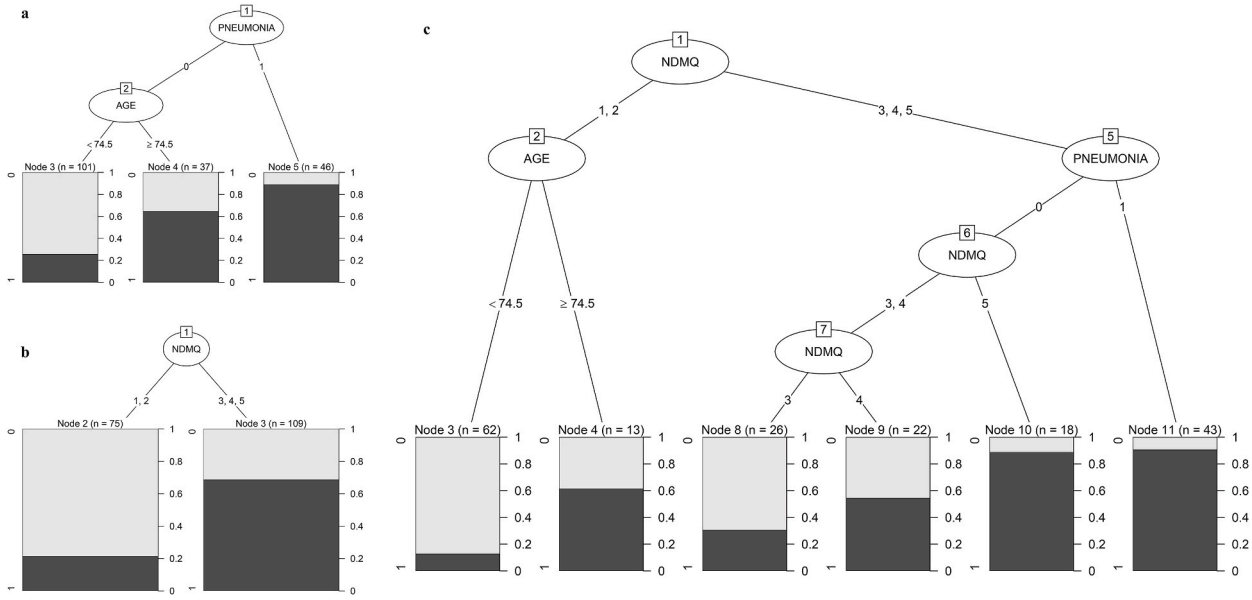

Fig. 3. The result of decision tree-classification tree three-step prediction model construction. a: Clinical information only. b: Laboratory test only. c: Screening characteristics included the characteristics as follows: NDM^Q, PNEUMONIA and AGE.

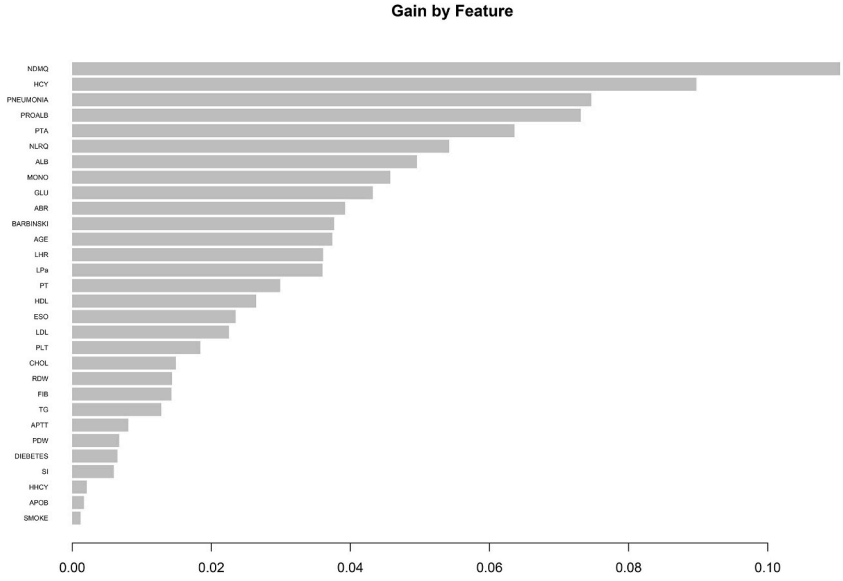

Fig. 4. Results of feature screening by Xgboost.

3.6. Comparison results of the ANN modeling

The characteristics in the ANN were that selected from the above four prediction models, which were as follows:
LASSO cross-validation regression: AGE, PNEUMONIA, BARBINSKI, PTA, HCY, PROALB, APOB, and NDM^Q (Fig. 6a);
Decision tree-classification tree: NDM^Q, PNEUMONIA, and AGE (Fig. 6b);
Xgboost: NDM^Q, HCY, PNEUMONIA, PROALB, PTA, and NLR^Q (Fig. 6c);
RF: NDM^Q, NLR^Q, PROALB, MONO, PT, HCY, and PNEUMONIA (Fig. 6d).

After the prediction model was built through the ANN, the confusion matrix was calculated separately (Table 2), so as to calculate the accuracy, sensitivity, specificity, PPV and NPV. The results showed that the accuracy, sensitivity, specificity, PPV and NPV for the LASSO-ANN model were 0.667, 0.567, 0.729, 0.567, and 0.729; for the decision tree-ANN model were 0.679, 0.583, 0.762, 0.677 and 0.681; for the Xgboost-ANN model were 0.577, 0.500, 0.658, 0.606 and 0.556; for the RF-ANN model were 0.718, 0.774, 0.680, 0.615,

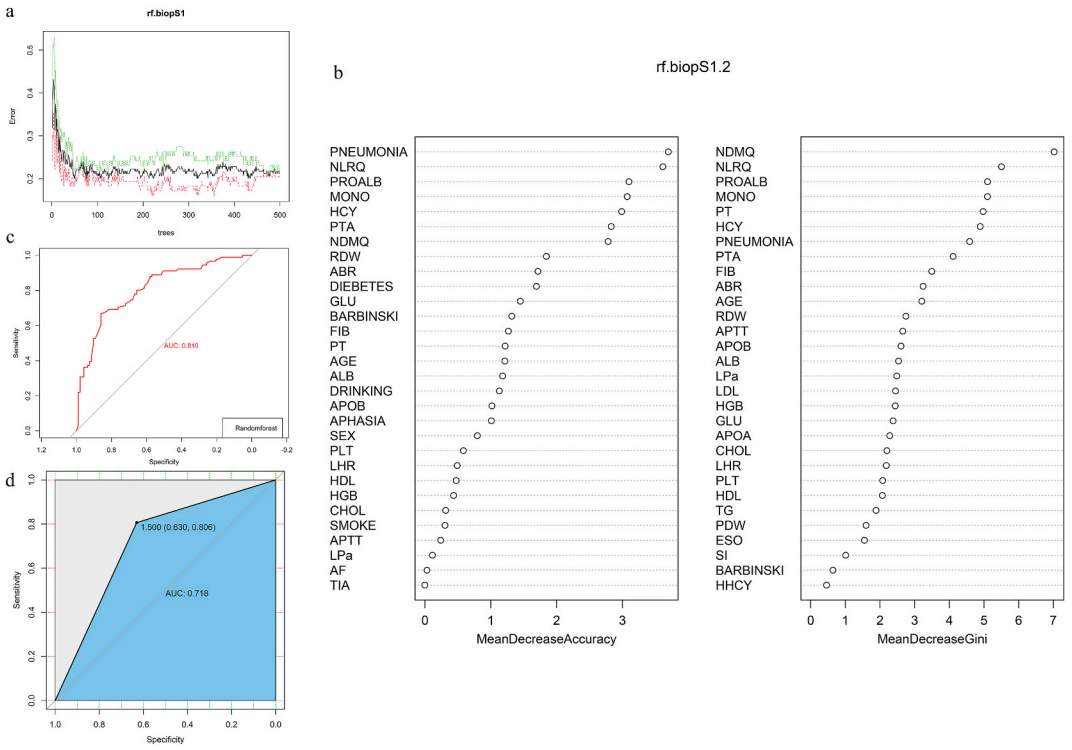

Fig. 5. Results of RF model construction. a: Figure of tree error graph in train set. b: The accuracy graph and Gini graph obtained after the RF model is set to the optimal biop. c: ROC curve of RF model in train set. d: ROC curve of RF model in test set.

and 0.821, respectively (Table 3).

3.7. External clinical validation of predicted models and comparison between models

The four external data patients were 659, 398, 1074, and 1239, respectively. The incidence of poor prognosis was 6.07 %, 6.43 %, 6.42 %, and 6.38 %, respectively. Only the accuracy of the model was shown (Table 4). The RF model performed best in external clinical validation datasets with accuracy of 0.646, 0.697, 0.695 and 0.713, respectively. Only the accuracy of LASSO cross-validated regression model was significantly correlated with sample size ($P = 0.026$). There was no significant correlation with sample size (Fig. 7a).

4. Discussion

The prediction models were constructed and compared by various ML in this study. The RF model achieves the best performance among medical prediction models with categorical variables as dependencies. NDM can be used in a variety of ML prediction models. In addition, if the sample size difference between groups is overly large, it is necessary to reduce the sample size difference between groups by random selection after clustering to obtain a more accurate model.

4.1. Features of the prediction models constructed by each ML method

Currently, no categorization algorithm has been found to have absolute an advantage for any industry or data. Extensive research has shown that many algorithms are similar in accuracy and need to be used on a case-by-case basis. To improve the rigor and objectivity, this study used supervised learning to construct four prediction models and compared the performance in the train and test sets of the ANN model. The results show that RF is able to screen variables better than the remaining three ML methods. RF is recommended for screening characteristics and modeling when the data volume is small and the dependent is categorical.

To improve the performance of predictive models, it is necessary to construct effective predictive models using reasonable statistical methods. The advantage of LASSO cross-validation regression is that the regression coefficients are clearly displayed, and it is convenient to implement the visualization of the model by plotting a nomogram, so that the model can be easily interpreted. However, the LASSO cross-validation regression model has certain limitations. (1) Some data will be lost when the continuous variable is converted to discrete values, and the trade-off of "regression unit" during the assignment increases some losses. Thus, the prediction of disease risk is only an approximation of the actual risk predicted in the complete model. (2) The regression coefficient, which is the

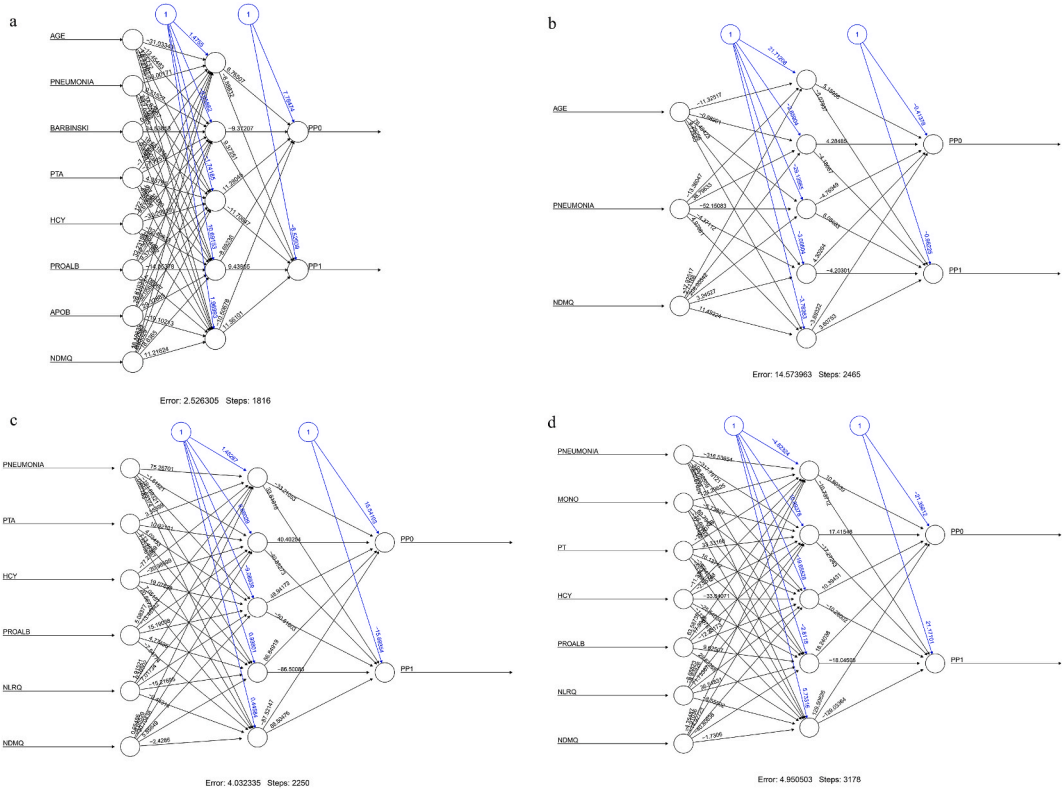

Fig. 6. a: Neural network model based on LASSO validation regression variable selection. b: Construct neural network model by filtering variables according to classification tree. c: Neural network was constructed according to Xgboost screening variables. d: Construction of neural network based on RF screening variables.

Table 4
The performance of LASSO cross-validation model, decision tree classification tree model and RF model in four external clinical datasets.

Datasets	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Sample Size	659	389	1074	1239
Incidence of poor prognosis	6.07	6.43	6.42	6.38
Accuracy				
LASSO cross validation	0.656	0.648	0.665	0.677
Decision tree-classification tree	0.590	0.573	0.577	0.600
Random Forest	0.646	0.697	0.695	0.713

basis of the LASSO cross-validation analysis, reflects the proportion of dependent variables that vary with the unit independent variable. However, the regression coefficients do not directly reflect the relationship between the derived variables. Therefore, the scoring form needs to compare the independent variables in order to better identify and analyze the relationship between them.

Compared to LASSO cross-validation regression, the decision tree-classification tree method has the advantage of rapid modeling speed and accurate conclusions. High degree of visualization and interpretation. The advantages of decision tree include: (1) the target dimension can be summarized according to the known dimension; (2) considering the influence of multiple known dimensions on the target dimensions, the robustness of the system is improved. (3) effectively inhibit the interference of noisy data; (4) effectively deal with the absence of some data; (5) it is more flexible to use as there is no need to make any prior assumptions; (6) effectively alleviate the problems in medical data acquisition, so as to improve the accuracy and reliability of the system. The disadvantage of decision tree is that it is easy to overfit in model construction. Fortunately, a reasonable pruning strategy can effectively balance the complexity of tree structure and the accuracy of the conclusion. In this study, we chose the post-prune of the decision tree algorithm, which controls the number of samples per node. Although this method will increase the overall calculation of the decision tree, it can produce a more accurate decision tree and facilitate the interpretation and utilization of the model. In order to discover the role of binary argument variables in data, decision tree needs to be carried out in steps when screening characteristics. As a result, improper use may drown some valuable characteristics or the issues they represent.

The RF algorithm with Bootstrap as the core constructed data sets through repeated random sampling and finally generated a forest composed of M decision trees. The results of these decision trees were integrated together for statistical analysis to effectively find out

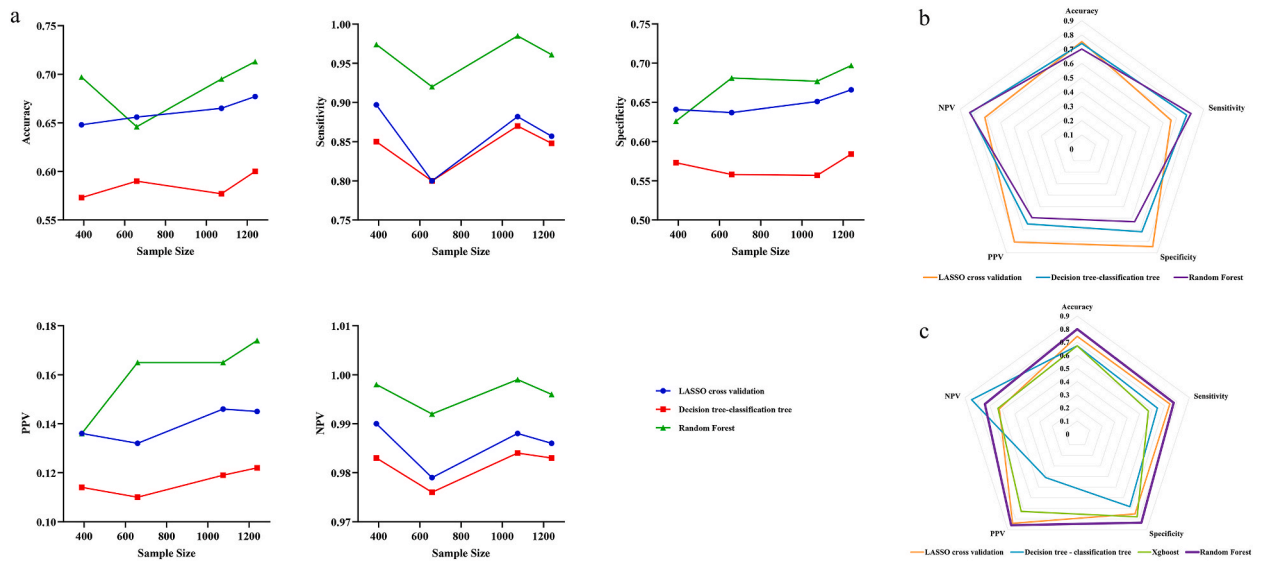

Fig. 7. a: The results were analyzed by plotting the performance of each model against the sample size. b: The performance comparison of LASSO cross-validation, decision tree-classification tree, RF in test set. c: Comparison of the ability of LASSO cross-validation, decision tree-classification tree, Xgboost, RF to filter variables.

the classification conclusions with the highest accuracy. By using Gini coefficient to measure the effect of current node division, this study adopts post-pruning strategy to control bifurcation nodes in RF algorithm, so as to make the tree structure more concise and efficient [12], and it can better reflect the effectiveness of node partitioning. Compared with decision tree, RF has the following advantages: (1) it can effectively solve local optimization and overfitting; (2) insensitivity to collinearity between dimensions during modeling; (3) the model has higher classification and prediction accuracy. The drawback of RF is that it is difficult to visualize the model. Currently, R is used to plot the RF Gini diagram to show the importance index of features in the model, the Gini coefficient. There is no relevant package to visualize the RF model in R, but the "graphviz" of Python algorithm can work at present. For medical workers, it is better to convert the model algorithms into software or web pages to facilitate the use of the model when running RF models.

ANN based algorithms are characterized by multiple inputs and a single output. In this study, the ANN model is constructed using the BP algorithm, which takes the data from the input layer, processes it through the intermediate layers, and finally obtains the predicted values in the output layer, thus enabling fast and accurate analysis of complex information. We perform extensive tests and comparisons on complex intermediate layer architectures via R to determine the best hidden layer architecture to improve accuracy. These hidden layers may contain one or more complex structures, each containing multiple nodes [13], as a result, we can better understand the complexity of the middle tier. The disadvantage of the ANN algorithm is the difficulty to analyze the fraction of features in the model from the model graph alone, and the model needs to be converted into software for ease of use in clinical applications. Some studies have found that the prediction accuracy of ML multilayer perceptron ANN model is better than that of multi-variate logistic regression model [14], which is slightly different from the results of this study. In this study, four ML screened characteristics were used to build ANN models, and the performance of the ANN model was worse than that of the respective ML models on the basis of selecting the same characteristics. The reason may be that the ANN is more suitable for datasets where independent variables are continuous variables. However, data standardization does not affect the ability of ML model to screen characteristics by using ANN for comparison.

At present, it has become a research trend to use multiple ML methods to construct and compare medical prediction models [8,15,16]. This study attempted to construct the short-term prediction model of poor prognosis of AIS based on ML methods, by collecting the clinical information of patients admitted to hospital and the results of the first laboratory examination after admission as the independent variables, and whether there was poor prognosis as the dependent variable. Each model could basically identify AIS patients who were likely to have poor prognosis over the next 3 months. The performance of the models is then compared and the RF model has the best performance for screening characteristics in this study. When the dependent variable was the classification variable in the medical data modeling, RF model could be the first choice to screen characteristics (Fig. 7b). According to the performance comparative analysis, RF model is the optimal model in this study, and can be selected in the construction of medical prediction models with dependent variables as categorical variables. In this study, due to the small amount of data after sample clustering, RF, as a major algorithm model, has been verified in medical data analysis of small data samples (Fig. 7c).

In terms of model visualization ability, the performance of LASSO cross-validation regression and decision tree-classification tree are better than RF when using only R. In the LASSO cross-validation regression model, the corrected C-statistic is 0.871 for the internal-validation and 0.825 for the external-validation, which is higher than the RF validation results. At the same time, the nomogram can be easily used by hospital staff. In the decision tree-classification tree model, the truncation values of characteristics and the probability of

poor prognosis were clearly shown in the model diagram, and the validation results were also good.

4.2. Comparison between this study and other prediction model studies

Among numerous prognostic prediction studies, only about 15 % of them focus on laboratory testing data. The predictive factors used to construct functional outcome prediction models are primarily blood glucose [7,17], neutrophil-to-lymphocyte ratio (NLR) [18], and C-reactive protein (CRP) [19,20]. Occasionally, studies include hemoglobin (HGB) [21], platelet count [18,22], serum triglycerides (TG) [23], and serum magnesium [24] as predictive factors. The NDM proposed in this study has only been reported in our previous study [11]. After summarizing the prediction models, it was found that in one case, the highest AUC (0.884) was established with NIHSS, RDW to platelet ratio, uric acid, 25 hydroxyvitamin D and angiopoietin-1 as characteristics [25], and others were all lower than 0.86. In this study, the LASSO cross-validation prediction model has strong accuracy, with C statistic up to 0.910 in the training set and 0.825 in the test set; the accuracy of RF prediction model in the training set was lower than that of LASSO cross-validation prediction model, but it had stronger sensitivity (0.806) and NPV (0.829). In terms of the AUC of the training set, the LASSO cross-validation prediction model in this study has been superior to the above model, and the data and information that can be easily obtained are used, which provides convenience for the use of clinical models.

4.3. Prediction value of NDM in this study

NDM^Q has predictive value in all ML models in this study. In the classification tree model, NDM^Q occupies the first tree node in the classification tree model established by screening feature variables, suggesting that NDM^Q plays an important role in the classification tree prediction model. Combined with clinical information, the advantages of test items as predictors of poor prognosis can be more fully utilized. NDM^Q combined with pneumonia may better predict the short-term prognosis of AIS patients. The RF prediction model can identify patients who are likely to have a poor prognosis in the short term, and with appropriate preventive measures or treatment, the occurrence of poor prognostic outcomes can be avoided.

Inflammation may cause damage to the blood-brain barrier, microvascular failure, brain edema and blood oxygen stress, which can directly cause neuronal cell death, leading to more serious brain damage [26], and the detection of AIS severity, prognosis assessment and other inflammatory indicators have become the main direction of scientific research. Neutrophils can damage host tissue and infiltrate damaged brain tissue shortly after the onset of AIS, leading to increased inflammation [27]. After ischemia and reperfusion, neutrophils accumulate in the meninges and perivascular spaces, eventually reaching the infarct parenchyma [28]. D-dimer is a special degradation product of cross-linked fibrin in the process of fibrinolysis. It is a sign of thrombin formation and fibrinolysis and has relatively stable characteristics [29]. Early D-dimer levels have been found to be an independent predictor of large vessel occlusion and may help prehospital patients better transfer to an appropriate stroke center [30]. In this study, these two factors were multiplied and combined to amplify their prediction effect on the short-term poor prognosis of AIS.

There are also studies on non-routine laboratory projects, including CCL11 [31], plasma copeptin [32], plasma Klotho [33], plasma RGM-A [34], plasma TAFI [35], plasma copeptin [32], serum irisin [36], serum Tau protein [37], serum tight-junction protein [38], serum retinoic acid [39], thrombospondin-1 [40], remnant lipoproteins [41], and Lipocalin-2 [42]. However, these non-routine laboratory tests face limitations in clinical application due to the difficulty in detection. Therefore, they are challenging to be used for building clinical rapid prediction models. These projects may require more complex and expensive laboratory techniques, as well as lacking uniform standards and ranges, which restrict their application in clinical settings. Therefore, in prognostic prediction research, the focus is mainly on routine laboratory tests to obtain more reliable and implementable prediction models. The independent variable for this study was selected as the laboratory data tested for the first time after admission. Combined with the medical history at admission, the data source was reliable and rapid, and the prediction model constructed provided speed advantages for the prediction of poor prognosis of AIS patients within 3 months.

NLR plays an essential role in LASSO cross-validation regression model and RF model. It is recommended that physicians pay more attention to NLR results in clinical applications. Studies have shown that women with AIS are more likely to have a poor short-term prognosis [43,44]. While in this study, no significant difference in short-term poor prognosis was found between the genders.

4.4. Predictive value of clinical information

Some of the accompanying illnesses or symptoms may be caused by the brain injury itself, by physical braking, or by AIS related treatments. These concomitant diseases or symptoms may hinder the recovery of the nervous system and have an important impact on the prognosis of patients with AIS [45]. Pneumonia accounted for a large proportion in this study, although as predictors they performed worse than NDM. It is worth mentioning that pneumonia is a prominent accompanying disease after AIS, accounting for 7–38 %. AIS triggers an inflammatory response in the brain that can lead to secondary brain damage and even infarct enlargement [46]. Multiple clinical studies have confirmed that pneumonia may be independently associated with poor prognosis and disability of AIS patients [47–49]. Previous studies have also shown that a variety of conditions are associated with the severity and outcome of AIS, such as myocardial infarction, stroke severity, pre-stroke disability, elevated intracranial pressure, pneumonia [44,50]. The odds of a poor prognosis were 5.08 times higher in those with pneumonia. The reason for these phenomena may be that the inflammatory immune response is more severe in AIS patients with pneumonia. This study showed that there was no significant correlation between previous smoking, SI and drinking history and short-term prognosis of AIS.

4.5. Limitations of this study

There are some limitations to this study. This study did not find a statistically significant difference in short-term poor outcomes between male and female patients, and a gender-based subgroup sensitivity analysis was not performed, so potential differences in gender-based outcomes could not be identified. The data in this study are from high latitudes in China. Regional bias may be present in the included subjects due to differences in food culture across different latitudes. Data from different latitudes can be included in subsequent studies for multi-center comprehensive analysis. Stroke severity-NIHSS score was not included in this study, because it was difficult to find in all medical histories and would have a large number of missing values once included, thus NDM was not compared with NIHSS score. This study has not verified the significant role of age in predicting AIS prognosis, and whether this is related to the trend towards younger age for AIS needs to be demonstrated further. Follow-up studies should be conducted with multi-center studies to establish age groups and to model.

5. Conclusion

By comparing various prediction models, this study concludes that the optimal choice of feature variables is the RF prediction model. In medical data analysis with small data samples and outcome as a categorical variable, RF can be used to construct prediction models for major algorithms. LASSO cross-validation regression and decision tree-classification tree models have better visualization ability. NDM, as the leader of independent predictor in this study, has predictive value in all ML models.

Funding

This research received the financial support by the Scientific Research Innovation Fund of the First Affiliated Hospital of HMU (2023M32) in conducting this research, analyzing the data, or preparing the manuscript for submission.

Ethics statement

This study was reviewed and approved by the Ethics Committee of the First Affiliated Hospital of HMU, with the approval number: 2020JS22.

Data availability statement

The datasets are not publicly available due to privacy or ethical restrictions, but they will be made available on request. Requests to access these datasets should be directed to Yinting XING, E-mail: 006883@hrbmu.edu.cn.

CRediT authorship contribution statement

Yinting Xing: Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition. **Yingyu Jin:** Resources, Investigation, Formal analysis, Data curation. **Yanhong Liu:** Writing - review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Xing Yinting directs the Innovation Fund of the First Affiliated Hospital of Harbin Medical University (2023M32) and is the lead author of the article.

Acknowledgments

Thank Mr. Zhiyuan Niu for his strong support in data collection, the laboratory department of the First Affiliated Hospital of HMU for providing research sites and experimental equipment, and Ms. Meixi Bao of the ethics committee of the First Affiliated Hospital of HMU for her guidance in ethics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e24232.

References

[1] P. Cui, L. McCullough, J. Hao, Brain to periphery in acute ischemic stroke: mechanisms and clinical significance, Front. Neuroendocrinol. (2021) 100932.
[2] L. Song, et al., A functional variant of the long noncoding RNA AL110200 is associated with the risk of ischaemic stroke recurrence, Eur. J. Neurol. 28 (8) (2021) 2708–2715.
[3] Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet Neurol. 20 (10) (2021) 795–820.
[4] A. Hidalgo, M. Casanova-Acebes, Dimensions of neutrophil life and fate, Semin. Immunol. (2021) 101506.
[5] J. Juega, et al., Monocyte-to-Lymphocyte Ratio in Clot Analysis as a Marker of Cardioembolic Stroke Etiology, Translational stroke research, 2021.
[6] Q. Meng, et al., Folic acid targets splenic extramedullary hemopoiesis to attenuate carbon black-induced coagulation-thrombosis potential, J. Hazard Mater. 424 (2021) 127354.
[7] L. Rinkel, et al., High admission glucose is associated with poor outcome after Endovascular treatment for ischemic stroke, Stroke 51 (11) (2020) 3215–3223.
[8] K. Lv, et al., Detection of diabetic patients in people with normal fasting glucose using machine learning 21 (1) (2023) 342.
[9] S. Desai, R. Jha, I. Linfante, Collateral circulation augmentation and neuroprotection as adjuvant to mechanical thrombectomy in acute ischemic stroke, Neurology 97 (2021) S178–S184.
[10] J. Heo, et al., Machine learning-based model for prediction of outcomes in acute stroke 50 (5) (2019) 1263–1265.
[11] Y. Xing, et al., Neutrophil count multiplied by D-dimer combined with pneumonia may better predict short-term outcomes in patients with acute ischemic stroke 17 (10) (2022) e0275350.
[12] L. Breiman, Random forests, [J] Machine learning 45 (2001) 5–32.
[13] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Internal Representations by Error Propagation, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
[14] L. Haifang, et al., Prediction of short-term outcome after subacute ischemic stroke using multiple layer perceptron neural network, Chin J Rehabil Theory Pract, Mar. 28 (2022), 3.
[15] Y. Zhuang, et al., Preoperative Prediction of Postoperative Infections Using Machine Learning and Electronic Health Record Data, 2023.
[16] Z. Wang, et al., A risk assessment framework for multidrug-resistant Staphylococcus aureus using machine learning and mass spectrometry technology 24 (6) (2023).
[17] P. Yang, et al., Admission fasting plasma glucose is an independent risk factor for 28-day mortality in patients with COVID-19, J. Med. Virol. 93 (4) (2021) 2168–2176.
[18] S.H. Lee, et al., The neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios predict reperfusion and prognosis after endovascular treatment of acute ischemic stroke, J Pers Med 11 (8) (2021).
[19] H. Naess, et al., C-reactive protein and homocysteine predict long-term mortality in young ischemic stroke patients, J. Stroke Cerebrovasc. Dis. 22 (8) (2013) e435–e440.
[20] I.U. Song, et al., Can high-sensitivity C-reactive protein and plasma homocysteine levels independently predict the prognosis of patients with functional disability after first-ever ischemic stroke? Eur. Neurol. 64 (5) (2010) 304–310.
[21] L. Li, et al., Impact of homocysteine levels on clinical outcome in patients with acute ischemic stroke receiving intravenous thrombolysis therapy, PeerJ 8 (2020) e9474.
[22] P. van der Meijden, J. Heemskerk, Platelet biology and functions: new concepts and clinical perspectives, Nat. Rev. Cardiol. 16 (3) (2019) 166–179.
[23] Y. Zhao, et al., Elevated triglyceride-glucose index predicts risk of incident ischaemic stroke: the Rural Chinese cohort study, Diabetes & metabolism 47 (4) (2021) 101246.
[24] D. Stewart, et al., Magnesium sulfate neither potentiates nor inhibits tissue plasminogen activator-induced thrombolysis, J. Thromb. Haemostasis 4 (7) (2006) 1575–1579.
[25] Z. Nannan, et al., Risk factors analysis and predictive model construction of poor early prognosis in acute ischemic stroke, ChinJClinRes, April2022 35 (No.4) (2022) 456–461.
[26] Y. Zhang, et al., The predictive role of systemic inflammation response index (SIRI) in the prognosis of stroke patients, Clin. Interv. Aging 16 (2021) 1997–2007.
[27] S. Kim, et al., Neutrophil extracellular trap induced by HMGB1 exacerbates damages in the ischemic brain, Acta neuropathologica communications 7 (1) (2019) 94.
[28] A. Otxoa-de-Amezaga, et al., Location of neutrophils in different compartments of the damaged mouse brain after severe ischemia/reperfusion, Stroke 50 (6) (2019) 1548–1557.
[29] G.D. Lowe, Fibrin D-dimer and cardiovascular risk, Semin. Vasc. Med. 5 (4) (2005) 387–398.
[30] A. Ramos-Pachón, et al., D-dimer as predictor of large vessel occlusion in acute ischemic stroke, Stroke 52 (3) (2021) 852–858.
[31] M. Roy-O'Reilly, et al., CCL11 (Eotaxin-1) levels predict long-term functional outcomes in patients following ischemic stroke, Transl Stroke Res 8 (6) (2017) 578–584.
[32] C.W. Wang, et al., Plasma levels of copeptin predict 1-year mortality in patients with acute ischemic stroke, Neuroreport 25 (18) (2014) 1447–1452.
[33] J.B. Lee, et al., Plasma Klotho concentrations predict functional outcome at three months after acute ischemic stroke patients, Ann. Med. 51 (3–4) (2019) 262–269.
[34] J. Zhong, et al., Reduced plasma levels of RGM-A predict stroke-associated pneumonia in patients with acute ischemic stroke: a prospective clinical study, Front. Neurol. 13 (2022) 949515.
[35] K. Jood, et al., Convalescent plasma levels of TAFI activation peptide predict death and recurrent vascular events in ischemic stroke survivors, J. Thromb. Haemostasis 10 (4) (2012) 725–727.
[36] H. Wu, et al., Serum levels of irisin predict short-term outcomes in ischemic stroke, Cytokine 122 (2019) 154303.
[37] J. Bielewicz, et al., Does serum Tau protein predict the outcome of patients with ischemic stroke? J. Mol. Neurosci. 43 (3) (2011) 241–245.
[38] R. Kazmierski, et al., Serum tight-junction proteins predict hemorrhagic transformation in ischemic stroke patients, Neurology 79 (16) (2012) 1677–1685.
[39] M. Xu, et al., Decreased serum retinoic acid may predict poor outcome in ischemic stroke patients, Neuropsychiatr Dis Treat 16 (2020) 1483–1491.
[40] M. Al Qawasmeh, A. Alhusban, F. Alfwaress, An evaluation of the ability of thrombospondin-1 to predict stroke outcomes and mortality after ischemic stroke, Int. J. Neurosci. (2020) 1–4.
[41] T. Nakamura, et al., High serum levels of remnant lipoproteins predict ischemic stroke in patients with metabolic syndrome and mild carotid atherosclerosis, Atherosclerosis 202 (1) (2009) 234–240.
[42] S. Hochmeister, et al., Lipocalin-2 as an infection-related biomarker to predict clinical outcome in ischemic stroke, PLoS One 11 (5) (2016) e0154797.
[43] H. Gu, et al., Sex differences in vascular risk factors, in-hospital management, and outcomes of patients with acute ischemic stroke in China, Eur. J. Neurol. 29 (1) (2021) 188–198.
[44] H. Koennecke, et al., Factors influencing in-hospital mortality and morbidity in patients treated on a stroke unit, Neurology 77 (10) (2011) 965–972.
[45] S. Kumar, M.H. Selim, L.R. Caplan, Medical complications after stroke, Lancet Neurol. 9 (1) (2010) 105–118.
[46] D. Ghelani, et al., Ischemic Stroke and Infection: a brief update on mechanisms and potential therapies, Biochem. Pharmacol. (2021) 114768.
[47] B. Hotter, et al., External validation of five scores to predict stroke-associated pneumonia and the role of selected blood biomarkers, Stroke 52 (1) (2021) 325–330.
[48] K. Nam, et al., High neutrophil-to-lymphocyte ratio predicts stroke-associated pneumonia, Stroke 49 (8) (2018) 1886–1892.
[49] R. Ji, et al., Novel risk score to predict pneumonia after acute ischemic stroke, Stroke 44 (5) (2013) 1303–1309.
[50] H. Duan, et al., Myocardial Infarction Is Associated with Increased Stroke Severity, In-Hospital Mortality, and Complications: Insights from China Stroke Center Alliance Registries, Journal of the American Heart Association, 2021 e021602.