Prediction of post-stroke epilepsy using machine learning method

Anjiao Peng
Department of Neurology and National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

Xiaorong Yang
Outpatient Department, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

Zhining Wen
College of Chemistry, Sichuan University, Chengdu, Sichuan, 610064, China

Wanling Li
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Yusha Tang
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Wanlin Lai
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Xiangmiao Qiu
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Lin Zhang
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Shixu He
Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

Lei Chen (leilei_25@126.com)
Sichuan University West China Hospital https://orcid.org/0000-0001-5263-5540

Research

Keywords: Post-stroke epilepsy, machine learning, support vector machine, random forest

Posted Date: April 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24983/v1

Prediction of post-stroke epilepsy using machine learning method

Anjiao Peng[1,†], Xiaorong Yang[2,†], Zhining Wen[3,4], Wanling Li[5], Yusha Tang[5], Wanlin Lai[5], Xiangmiao Qiu[5], Lin Zhang[5], Shixu He[5], Lei Chen[1,4,*]

[1] Department of Neurology and National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

[2] Outpatient Department, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

[3] College of Chemistry, Sichuan University, Chengdu, Sichuan, 610064, China.

[4] Medical Big Data Center, Sichuan University, Chengdu, Sichuan, 610064, China.

[5] Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China.

[†] These authors contribute equally to this work.

[*] Corresponding Author: Lei Chen, Department of Neurology, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang, Chengdu, Sichuan 610041, China. E-mail: leilei_25@126.com.

Abstract

Background: Stroke is one of the most important causes of epilepsy and we aimed to find if it is possible to predict patients with high risk of developing post-stroke epilepsy (PSE) at the time of discharge using machine learning methods.

Methods: Patients with stroke were enrolled and followed at least one year. Machine learning methods including support vector machine (SVM), random forest (RF) and logistic regression (LR) were used to learn the data.

Results: A total of 2730 patients with cerebral infarction and 844 patients with cerebral hemorrhage were enrolled and the risk of PSE was 2.8% after cerebral infarction and 7.8% after cerebral hemorrhage in one year. Machine learning methods showed good performance in predicting PSE. The area under the receiver operating characteristic curve (AUC) for SVM and RF in predicting PSE after cerebral infarction was close to 1 and it was 0.92 for LR. When predicting PSE after cerebral hemorrhage, the performance of SVM was best with AUC being close to 1, followed by RF (AUC = 0.99) and LR (AUC = 0.85) .

Conclusion: Machine learning methods could be used to predict patients with high risk of developing PSE, which will help to stratify patients with high risk and start treatment earlier. Nevertheless, more work is needed before the application of thus intelligent predictive model in clinical practice.

Key words*: Post-stroke epilepsy; machine learning; support vector machine; random forest*

Background

Stroke has been recognized as one of the most important causes of epilepsy, especially in the elderly [1, 2]. Previous studies showed that about half of the patients over 60 years old with acquired epilepsy were caused by stroke [3, 4]. With the increasing number of aging population, the number of patients with post-stroke epilepsy (PSE) will continue to increase. PSE significantly reduced the quality of life of stroke patients, and increased the economic and psychological burden of the patients and their families [5]. Therefore, the prediction of patients with high risk of PSE is very important [2, 6].

Seizure can occur in the acute phase of, or months or years after the stroke. Seizures occur within 7 days after stroke are known as early onset seizures and are called late onset seizures or unprovoked seizures when occur 7 days after stroke [7]. According to the definition of the International League Against Epilepsy (ILAE), the late onset seizure is also known as PSE [8]. The risk of PSE is 2.6% to 9.5% in 5 years after cerebral infarction, which is 10 to 30 times that of the general population [9-13]. And it could be up to 11.8% after cerebral hemorrhage [14].

Although it is recognized that early onset seizure, cortical involvement, and the severity of symptoms are closely related to PSE [14], other clinical features and laboratory findings have not been fully explored. Few studies have demonstrated the feasibility of developing prediction models of PSE using traditional analyses, but their clinical application were limited due to low sensitivity and specificity [2, 6, 14]. Different from traditional analysis methods, machine learning method is superior in data mining and has become a hotspot in medical research [15]. Its powerful classification and predictive capabilities have been certificated by increasing studies [15].

In this study, we aimed to discover whether it is plausible to predict which patient would develop PSE based on their clinical features and laboratory findings. Three machine learning algorithms including

support vector machine (SVM), random forest (RF) and logistic regression (LR) were used to learn the data and build the predictive model.

Materials and methods

Date source and study population

This study was approved by the biomedical ethics committee of West China Hospital of Sichuan University. All subjects agreed to participate in the project and signed the informed consent. Patients were included if they were treated in Western hospital, Sichuan university from 2010 to 2017, diagnosed as "cerebral infarction" or "cerebral hemorrhage" and older than 16 years. Patients were excluded if they had (1) a previous history of epilepsy; (2) a previous history of cerebral hemorrhage or cerebral infarction; (3) intracranial tumor, infection, trauma or operation; (4) cerebral infarction caused by venous sinus thrombosis; (5) too much missing information; (6) cerebral infarction or cerebral hemorrhage occurred during follow-up; (7) died in hospital or within one year after discharge; (8) a follow-up period less than 12 months.

Data extraction

Information including demographic characteristics (gender, age at stroke etc.), clinical features (early onset seizure, antiepileptic drugs, risk factors of stroke, severity of stroke etc.), examinations results (biochemical and imaging examination findings) were systematically extracted. Seizures and antiepileptic drugs after discharge were obtained through clinical visit and (or) telephone follow-up based on a structured form. Seizures after 7 days after stroke were diagnosed as PSE.

Missing data handling

Missing data is an inevitable problem and removing all subjects or variables with missing data would

loss lots of information and reduce the sample size. The following process of missing data handling was done before machine leaning algorithms were applied: a) variables with too much missing data and classified as possibly unimportant according to the clinical experience and previous research results were removed; b) subjects with too much missing records were also removed; c) the missing data then be supplemented by median (quantitative data) or mode (categorical data).

Features selection

The establishment of models largely depends on the correct selection of features. More features included would cause much more noise and result in overfitting of the models. But the performance of model would be affected if not enough features were included [16]. In this study, Univariate feature selection method was used to select features, which calculated the score of each feature and the P value of each variable in a scoring function. Features with P value less than 0.05 were then put into the machine learning algorithm.

Class unbalance

Our data showed that the number of patients who developed PSE was far less than those who did not, the resulting classification are characterized by an unbalanced distribution of the class variable. It is impossible to ignorance the class unbalance issue since all subjects would be likely automatically classified into the majority class by machine learning methods. In this study, we used a common strategy in machine learning, the Synthetic Minority Oversampling Technology (SMOTE), to handle the class unbalance issue [17]. Simply speaking, new samples were synthesized based on the character of subject and added to the data set of the minority class.

Models built

The samples were randomly divided into training set (70%) and testing set (30%). RF, SVM and LR

were used to learn the data and build the prediction models. Decision trees which were built by randomly selecting subset features to best separate the data into the expected outputs. The forest of decision trees was then generated and to form the final output [18]. SVM maps the data to the high-dimensional space through kernel function to maximally separate the clusters of data which were not separable in low-dimensional space [19].

Assessment of the performance of predictive models

The final step was to assess the performance of the predictive models. the accuracy of different models was evaluated by receiver operating characteristic curve (ROC). The closer the area under the ROC curve (AUC) is to 1, the more accurate the model is. Other indicators including sensitivity, specificity, positive predictive value and negative predictive value were also measured for the models.

Results

Clinical features and examination results of patients

A total of 4853 patients with acute cerebral infarction and 1626 patients with cerebral hemorrhage were enrolled. According to the exclusion criteria, 2730 patients with cerebral infarction and 844 patients with cerebral hemorrhage were included. The flow chart of patient inclusion and exclusion was shown in figure 1. The average follow-up period was 28.0 ± 17.9 months for patients with cerebral infarction and 77 (2.8%) patients developed PSE within one year. For patients with cerebral hemorrhage, the mean follow-up period was 25.8 ± 15.9 months and 66 (7.8%) patients developed PSE within one year. Clinical features and examination results were summarized in (Table 1).

Features related to PSE

A total of 35 variables was found to be related to the PSE after cerebral infarction (Fig. 2A). The top

five were creatine kinase, hospitalization days, lactate dehydrogenase, early onset seizures and antiepileptic drugs used in acute phase of stoke. A total of 19 variables were found to be related to PSE after cerebral hemorrhage (Fig. 2B) and the top five were hospitalization days, uroleukocyte, frontal cerebral hemorrhage, alanine aminotransferase and early onset seizures.

Performance of different algorithms in predicting PSE

To assess the performance of different algorithms in predicting PSE, sensitivity, specificity, positive predictive value, negative predictive value and AUC were calculated (Table 2). The results showed that the performance of SVM and RF were better in predicting PSE after cerebral infarction with AUC being close to 1 in training set. The sensitivity, specificity positive predictive value and negative predictive value of them were also high. The performance of LR was a little poor with AUC being 0.92. In testing set, the performance of SVM and RF were also good, with AUC being close to 1, and the AUC of LR was 0.92 (Fig. 3A).

Similar to that in cerebral infarction, SVM and RF also performed good in predicting PSE after cerebral hemorrhage. In training set, SVM achieved the highest AUC which was close to 1, followed by RF, with AUC being 0.99. The the AUC of LR (0.85) was slightly lower. In testing set, the performance of RF (AUC = 0.98) and SVM (AUC = 0.97) were also better than LR (AUC = 0.86) (Fig. 3B). The sensitivity and specificity of RF and SVM were also higher than that of LR (Table 2).

Discussion

The results of this study showed that the risk of PSE was close to 3% after cerebral infarction and 8% after cerebral hemorrhage within one year, which was similar to previous studies [6, 14, 20]. The first year was the peak time for seizure relapse after stroke and we considered this temporal threshold in this

study [2]. Our model is designed to be used at the time of discharge and it showed that it is feasible to stratify patients with high risk of developing PSE using machine learning method based on clinical information and examination findings. Considering the differences of disease characteristics between cerebral infarction and cerebral hemorrhage, we analyzed them separately.

Hitherto, a few PSE prediction tools have been developed using traditional analysis, but their clinical application were limited because of complexity of operation, low sensitivity and poor specificity [2, 6, 14, 21]. In this study, we used intelligent analyses to predict PSE for the first time and it showed that the performance of SVM and RF were the best with high sensitivity, specificity and AUC being close to 1. SVM is now widely used in classification since the kernel used in SVM model is a shortcut to accelerate the learning process and greatly improve the accuracy of the model [19]. RF is also widely used because it is easy to calculate, has high accuracy, can process large data sets and does not need to reduce the dimension of high-dimensional data sets [18]. However, we should also know that SVM and RF could be heavily influenced by the unbalance class issue [22]. The unbalance class issue was handled using SMOTE methods in this study, which may reduce the differences among synthesized samples and results in better performance of SVM and RF. Another shortage of these two algorithms is that the calculation process is a “black box”, which could not be easily explained and understand by clinicians. However, the value of SVM and RF in predicting PSE was undeniable considering their superior performance evaluated by sensitivity, specificity, positive predictive value, negative predictive value and AUC. Larger longitudinal studies were needed to test the application of these models in clinical experience.

Importantly, The results of this study showed that neither endovascular thrombectomy nor thrombolysis with recombinant tissue plasminogen activator would increase the risk of PSE and it is always been the

focus of clinicians' attention whether reperfusion treatment would increase the risk of PSE in cerebral infarction [2, 23, 24]. Similar with previous studies, we found that early onset seizure, symptom severity (which was assessed by NIHSS score at admission, length of stay in hospital, massive cerebral infarction and haemorrhagic transformation) and cortical involvement were related to PSE [2, 6, 13, 14]. What is more, many laboratory findings like urine leukocytes, uric acid, alanine aminotransferase, creatine kinase and lactate dehydrogenase were also found to be associated with PSE. Uric acid is now believed to be inflammatory and has been confirmed by previous clinical and basic researches that both increase and decrease of uric acid level could lead to an increased incidence of PSE [25, 26]. But the association between other laboratory findings and PSE need further research.

There are some limitations in this study. First, since only a relatively minority of patients developed PSE, which resulted in significant class unbalance, we used the SMOTE, a method widely used in dealing such issue to handle the unbalance class issue in artificial analyses. Which may lead to better performance of predictive models than they actually do. Second, due to the limited amount of data, we only constructed models to predict PSE one year after stroke, which may limit its clinical use. Finally, although the predictive models all showed good performance, similar to previous study, we can only discuss the possibility and accuracy of intelligent analyses in predicting PSE. The use of such models in clinical practice still has a long way to go.

Conclusion

This study demonstrated that the risk of PSE is about 2.8% after cerebral infarction and 7.8% after cerebral hemorrhage in one year. Lots of new risk factors were found to be related to PSE and based on these variables. We successfully constructed predictive models and RF and SVM showed better

performance than LR in predicting PSE both in patients with cerebral infarction and cerebral hemorrhage.

Abbreviations

AUC: area under the ROC curve; LR: logistic regression; PSE: post-stroke epilepsy; RF: random forest; ROC: receiver operating characteristic curve; SMOTE: Synthetic Minority Oversampling Technology; SVM: support vector machine.

Acknowledgements

Not applicable.

Authors' contributions

Peng, Yang and Chen conceived and designed the study. Li, Tang, Lai and Qiu performed patients follow-up and collected the data. Wen performed data analysis and interpretation. Peng and Yang drafted the manuscript. Zhang, He and Chen revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University (No.Z2018B24).

Availability of data and materials

The dataset analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This study was approved by the biomedical ethics committee of West China Hospital of Sichuan University. All subjects agreed to participate in the project and signed the informed consent.

Consent for publication

Not applicable.

Competing interests

Authors declare that they have no disclosures to report.

References:

1. Jin J, Chen R, Xiao Z. Post-epilepsy stroke: A review. Expert Review of Neurotherapeutics. 2016;16:341-349.
2. Galovic M, Döhler N, Erdélyi-Canavese B, Felbecker A, Siebel P, Conrad J*, et al*. Prediction of late seizures after ischaemic stroke with a novel prognostic model (the SeLECT score): a multivariable prediction model development and validation study. The Lancet. Neurology. 2018;17:143-152.
3. Brodie MJ, Kwan P. Epilepsy in elderly people. BMJ (Clinical Research Ed.). 2005;331:1317-1322.

4. Camilo O, Goldstein LB. Seizures and epilepsy after ischemic stroke. Stroke. 2004; 35:1769-1775.

5. Winter Y, Daneshkhah N, Galland N, Kotulla I, Krüger A, Groppa S. Health-related quality of life in patients with poststroke epilepsy. Epilepsy Behav. 2018;80:303-306.

6. Chi N, Kuan Y, Huang Y, Chan L, Hu C, Liu H*, et al*. Development and validation of risk score to estimate 1-year late poststroke epilepsy risk in ischemic stroke patients. Clinical Epidemiology. 2018;10:1001-1011.

7. Leung T, Leung H, Soo YOY, Mok VCT, Wong KS. The prognosis of acute symptomatic seizures after ischaemic stroke. Journal of Neurology, Neurosurgery, and Psychiatry. 2017;88:86-94.

8. Fisher RS (2017). The New Classification of Seizures by the International League Against Epilepsy 2017. Current Neurology and Neuroscience Reports, 17:48.

9. Adelöw C, Andersson T, Ahlbom A, Tomson T. Prior hospitalization for stroke, diabetes, myocardial infarction, and subsequent risk of unprovoked seizures. Epilepsia. 2011;52:301-307.

10. Burn J, Dennis M, Bamford J, Sandercock P, Wade D, Warlow C. Epileptic seizures after a first stroke: the Oxfordshire Community Stroke Project. BMJ (Clinical Research Ed.). 1997;315:1582-1587.

11. Chen T, Chen Y, Cheng P, Lai C. The incidence rate of post-stroke epilepsy: a 5-year follow-up study in Taiwan. Epilepsy Research. 2012;102:188-194.

12. Graham NSN, Crichton S, Koutroumanidis M, Wolfe CDA, Rudd AG. Incidence and associations of poststroke epilepsy: the prospective South London Stroke Register. Stroke. 2013; 44:605-611.

13. Roivainen R, Haapaniemi E, Putaala J, Kaste M, Tatlisumak T. Young adult ischaemic stroke related acute symptomatic and late seizures: risk factors. European Journal of Neurology.

2013;20:1247-1255.

14. Haapaniemi E, Strbian D, Rossi C, Putaala J, Sipi T, Mustanoja S*, et al*. The CAVE score for predicting late seizures after intracerebral hemorrhage. Stroke. 2014;45:1971-1976.

15. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (New York, N.Y.). 2015;349:255-260.

16. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH*, et al*. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal. 2017;38:500-507.

17. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106.

18. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Studies in Health Technology and Informatics. 2015;216:40-44.

19. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics (Oxford, England). 2000;16:906-914.

20. So EL, Annegers JF, Hauser WA, O'Brien PC, Whisnant JP. Population-based study of seizure disorders after cerebral infarction. Neurology. 1996;46:350-355.

21. Strzelczyk A, Haag A, Raupach H, Herrendorf G, Hamer HM, Rosenow F. Prospective evaluation of a post-stroke epilepsy risk scale. Journal of Neurology. 2010;257:1322-1326.

22. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V*, et al*. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol. 2018;12:295-302.

23. De Reuck J, Van Maele G. Acute ischemic stroke treatment and the occurrence of seizures.

Clinical Neurology and Neurosurgery. 2010;112:328-331.

24. Alvarez V, Rossetti AO, Papavasileiou V, Michel P. Acute seizures in acute ischemic stroke: does thrombolysis have a role to play? Journal of Neurology. 2013;260:55-61.

25. Wang D, Hu B, Dai Y, Sun J, Liu Z, Feng Y*, et al.* Serum Uric Acid Is Highly Associated with Epilepsy Secondary to Cerebral Infarction. Neurotoxicity Research. 2019;35:63-70.

26. Thyrion L, Raedt R, Portelli J, Van Loo P, Wadman WJ, Glorieux G*, et al.* Uric acid is released in the brain during seizure activity and increases severity of seizures in a mouse model for acute limbic seizures. Experimental Neurology. 2016;277:244-251.

Table 1. Basic demographic, clinical characteristics and laboratory results of patients with cerebral infarction (CI) and cerebral hemorrhage (CH).

	Features	Patients with CI	Patients with CH
Demographic features	Age (years)	61.7±14.5	53.0 ± 16.7
	Gender (male)	1706（62.5%）	552（65.4%）
	Marriage		
	Married	2498（91.5%）	750（88.9%）
	Other	232（8.5%）	94（11.1%）
	Smoking (yes)	1070（39.2%）	260（30.8%）
	Alcoholism		
	No	2027（69.2%）	593（70.3%）
	Occasionally	412（15.1%）	141（16.7%）
	Often	431（15.8%）	110（13.0%）
	Early onset seizure	60（2.2%）	53（6.3%）
	NIHSS score	4.5 ± 4.9	-
	Atherosclerotic type	1832（67.1%）	-
	Arteriolar occlusive type	535（19.6%）	-
	Cardiogenic embolism	139（5.1%）	-
	Other clear etiology	131（4.8%）	-
	Unknown cause	93（3.4%）	-
Clinical features	Conscious disturbance	224（8.2%）	325（38.5%）
	Hypertension	169（62.0%）	519（61.5%）
	Diabetes	81（29.7%）	89（10.6%）
	Pulmonary infection	396（14.5%）	177（21.0%）
	Electrolyte disturbance	41（1.5%）	72（8.5%）
	Cardiac insufficiency	109（4.0%）	19 （2.3%）
	Atrial fibrillation	408（14.9%）	21（2.5%）
	Hypoproteinemia	79（2.9%）	71（8.4%）

	Features	Patients with CI	Patients with CH
	Hyperlipidemia	49（11.8%）	22（2.6%）
	Renal insufficiency	68（2.5%）	38（4.5%）
	Hepatic insufficiency	87（3.2%）	26（3.1%）
Vital signs	Temperature (℃)	36.5 ± 0.3	36.6±0.1
	Pulse (times / minute)	93.1 ± 14.8	-
	Systolic pressure (mmHg)	138.4 ± 19.0	151.2 ± 29.1
	Diastolic pressure (mmHg)	87.2 ± 10.5	91.1 ± 18.0
Treatments	Antiepileptic therapy	63（2.3%）	-
	Intravenous thrombolysis	71（2.6%）	-
	Arterial thrombectomy	14（0.5%）	-
	Craniotomy	22（0.8%）	270（32.0%）
	Length of stay (days)	14.2 ± 18.4	16.9 ± 23.0
Laboratory findings	Hemoglobin (g / L)	133.7 ± 18.8	127.8 ± 21.3
	Leukocyte count (10 ^ 9 / L)	7.1 ± 2.4	9.1 ± 3.6
	Platelet count (10 ^ 9 / L)	183.5 ± 69.7	191.3 ± 86.4
	Eosinophils (10 ^ 9 / L)	0.2 ± 0.2	0.1 ± 0.2
	Basophils (10 ^ 9 / L)	0.03 ± 0.02	0.02 ± 0.02
	Neutrophils (10 ^ 9 / L)	4.9 ± 2.2	7.1 ± 3.5
	Lymphocytes (10 ^ 9 / L)	1.6 ± 0.6	1.3 ± 0.6
	Monocytes (10 ^ 9 / L)	0.4 ± 0.2	0.5 ± 0.3
	ANCA (positive)	718（26.3%）	-
	Serum CA-125 (U / ml)	20.2 ± 86.4	-
	Serum CA-153 (U / ml)	12.0 ± 9.8	-
	Serum CA-199 (U / ml)	15.6 ± 42.1	-
	CEA (ng/ml)	2.8 ± 6.9	-
	Urea (mmol / L)	5.8 ± 2.6	6.0 ± 3.5
	High density lipoprotein	1.2 ± 0.3	1.2 ± 0.4
	Low density lipoprotein (mmol / L)	2.3 ± 0.8	2.5 ± 0.7
	Total protein (g / L)	66.4 ± 5.6	66.5 ± 6.9

	Features	Patients with CI	Patients with CH
	Globulin (g / L)	26.7 ± 4.1	27.4 ± 5.3
	Albumin (g / L)	39.7 ± 4.2	39.0 ± 4.7
	Total bilirubin (μ mol / L)	13.7 ± 9.9	13.9 ± 8.3
	Direct bilirubin (μ mol / L)	5.0 ± 6.6	5.3 ± 4.3
	Cholesterol (mmol / L)	4.1 ± 1.1	4.3 ± 1.0
	Blood glucose (mmol / L)	6.5 ± 2.6	6.7 ± 2.3
	Aspartate aminotransferase (U / L)	27.5 ± 18.2	31.2±36.2
	Alanine aminotransferase (U / L)	28.0 ± 24.2	33.3 ± 39.6
	AST/ALT	1.2 ± 0.6	1.2 ± 0.6
	Indirect bilirubin (μ mol / L)	8.7 ± 4.7	8.5 ± 4.8
	Alkaline phosphatase (U / L)	81.9 ± 29.3	89.3 ± 44.1
	Glutamyltranspeptidase (U / L)	40.9 ± 56.5	53.2 ± 70.1
	Creatine kinase (U / L)	101.8 ± 236.6	191.8 ± 455.5
	Lactate dehydrogenase (U / L)	195.0 ± 67.7	217.1 ± 97.7
	Creatinine (umol / L)	80.1 ± 35.7	87.4 ± 107.7
	Uric acid (umol / L)	308.1 ± 96.6	265.5 ± 115.2
	Sodium (mmol / L)	141.2 ± 3.3	140.6 ± 4.6
	Calcium (mmol / L)	2.2 ± 0.1	2.2 ± 0.1
	Magnesium (mmol / L)	0.9 ± 0.1	0.9 ± 0.1
	Phosphorus (mmol / L)	1.1 ± 0.2	1.1 ± 0.3
	Chlorine (mmol / L)	104.5 ± 4.1	103.6 ± 5.2
	Potassium (mmol / L)	3.9 ± 0.4	3.9 ± 0.5
	Urinary leukocytes (PCS / HP)	11.1 ± 124.1	7.1 ± 43.5
Imaging features	Cerebral hernia	16（0.6%）	19（2.3%）
	Supratentorial CI/CH	2484（91.1%）	652（77.3%）
	Infratentorial CI/CH	349（12.8%）	116（13.7%）

Features	Patients with CI	Patients with CH
Temporal lobe involvement	704（25.8%）	149（17.7%）
Frontal lobe involvement	720（26.4%）	115（13.6%）
Insular lobe involvement	229（8.4%）	15（1.8%）
Parietal involvement	595（21.8%）	125（14.8%）
Occipital lobe involvement	393（14.4%）	71（8.4%）
Basal ganglia infarction	1250（45.8%）	362（42.9%）
Massive cerebral infarction	150（5.5%）	-
Hemorrhage transformation	292（10.7%）	-
With subarachnoid hemorrhage	-	117（13.9%）
With intraventricular hemorrhage	-	226（26.8%）
Secondary cerebral infarction	-	34（4.0%）

Data are expressed as mean±SD or frequency and percentage, as appropriate

AST: Aspartate aminotransferase; ALT:Alanine aminotransferase; CH: cerebral hemorrhage; CI: cerebral infarction.

Table 2. The performance of different algorithms in predicting PSE.

PSE after cerebral infarction (training set)					
	Sensitivity	Specificity	PPV	NPV	AUC
RF	99.46%	98.03%	98.01%	99.46%	1
SVM	100.00%	100.00%	100.00%	100.00%	1
LR	88.12%	81.69%	82.46%	87.56%	0.92

PSE after cerebral infarction (testing set)					
	Sensitivity	Specificity	PPV	NPV	AUC
RF	99.02%	95.35%	95.74%	98.93%	1
SVM	99.27%	97.29%	97.48%	99.21%	1
LR	86.80%	83.46%	84.73%	85.68%	0.92

PSE after cerebral hemorrhage (training set)					
	Sensitivity	Specificity	PPV	NPV	AUC
RF	96.72%	94.64%	94.81%	96.60%	0.99
SVM	100.00%	99.82%	99.82%	100.00%	1
LR	75.73%	77.82%	77.57%	75.99%	0.85

PSE after cerebral hemorrhage (testing set)					
	Sensitivity	Specificity	PPV	NPV	AUC

RF	92.17%	92.41%	92.17%	92.41%	0.98
SVM	87.83%	94.51%	93.95%	88.89%	0.97
LR	75.65%	80.59%	79.09%	77.33%	0.86

AUC = area under the ROC curve; LR = logistic regression; NPV = negative predictive value; PPV = positive predictive value; RF = random forest; SVM = support vector machine

Fig.1. The flow chart of patient inclusion and exclusion.

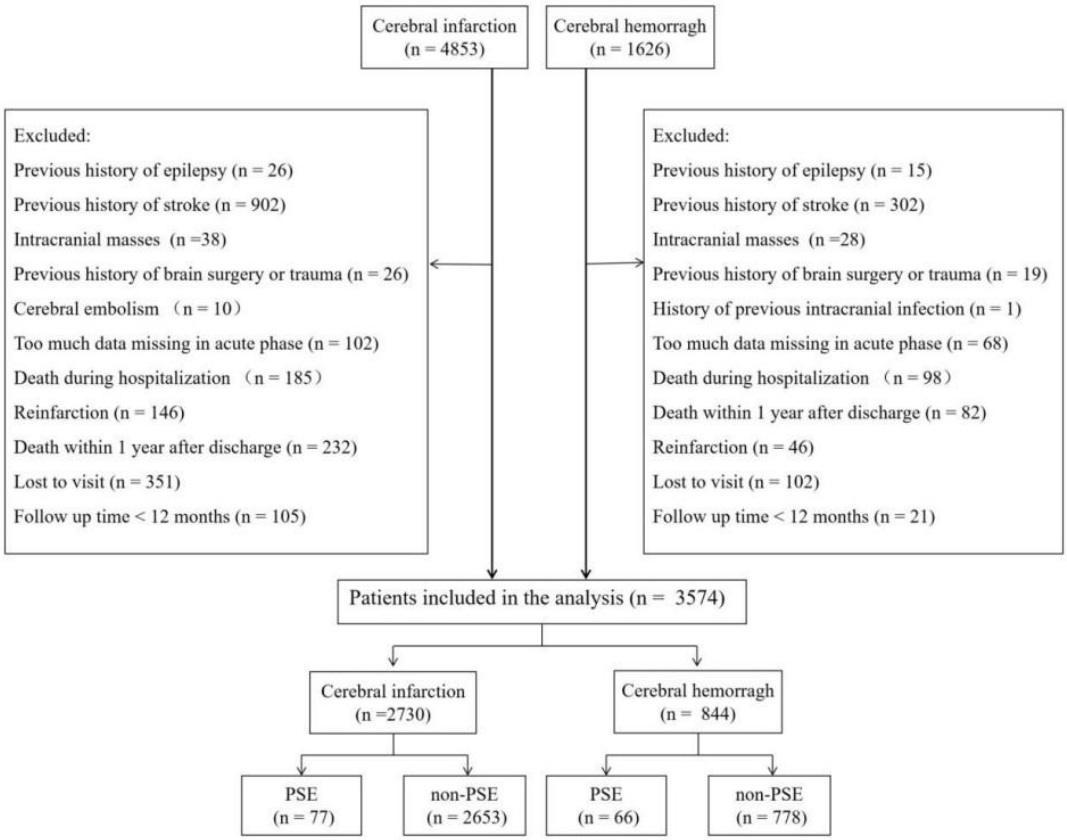
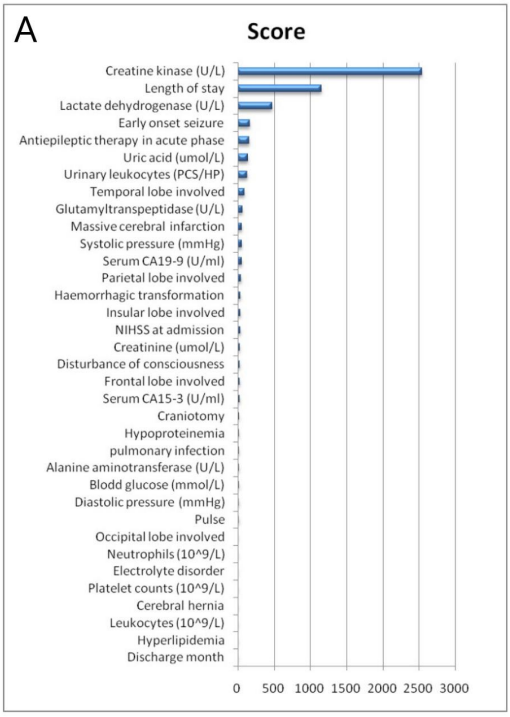


Fig.2. Characteristic variables related to PSE after cerebral infarction (A) and cerebral hemorrhage (B).



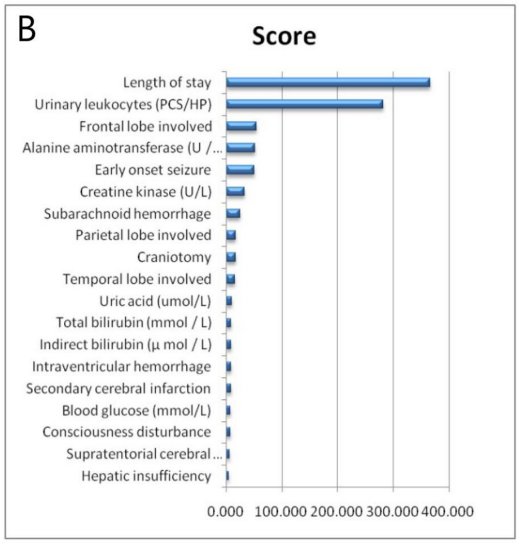
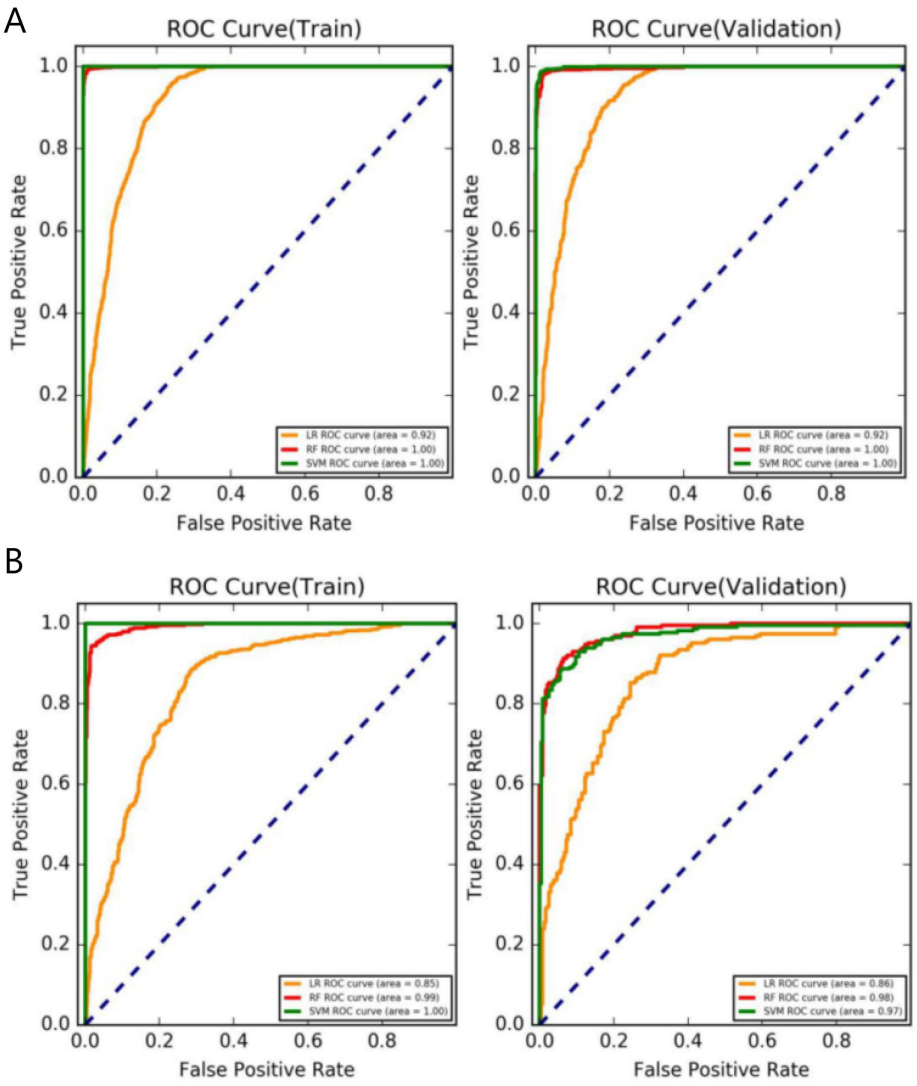


Fig.3. Accuracy of different algorithms in predicting PSE after cerebral infarction (A) and cerebral hemorrhage (B).



Figures

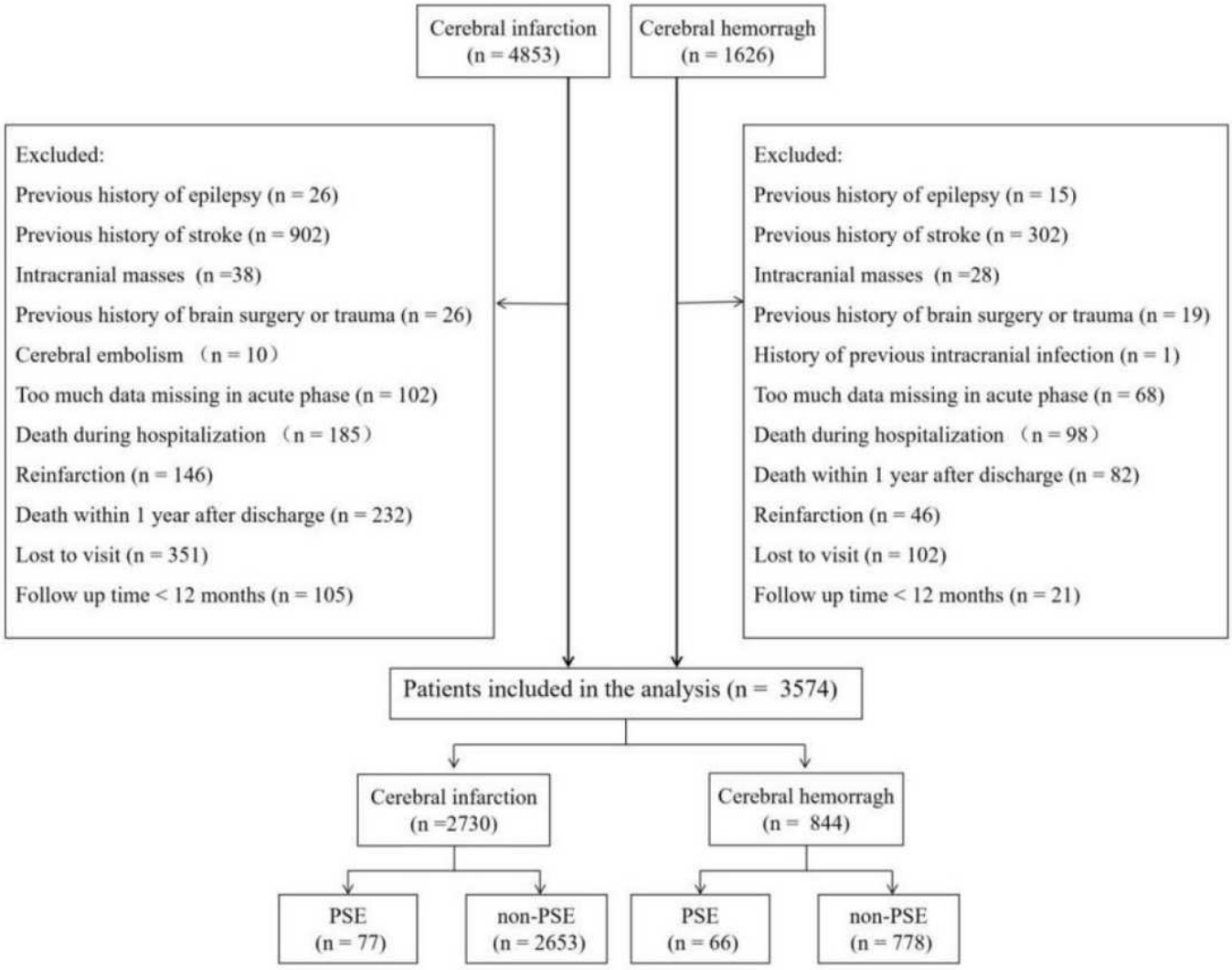


Figure 1

The flow chart of patient inclusion and exclusion.

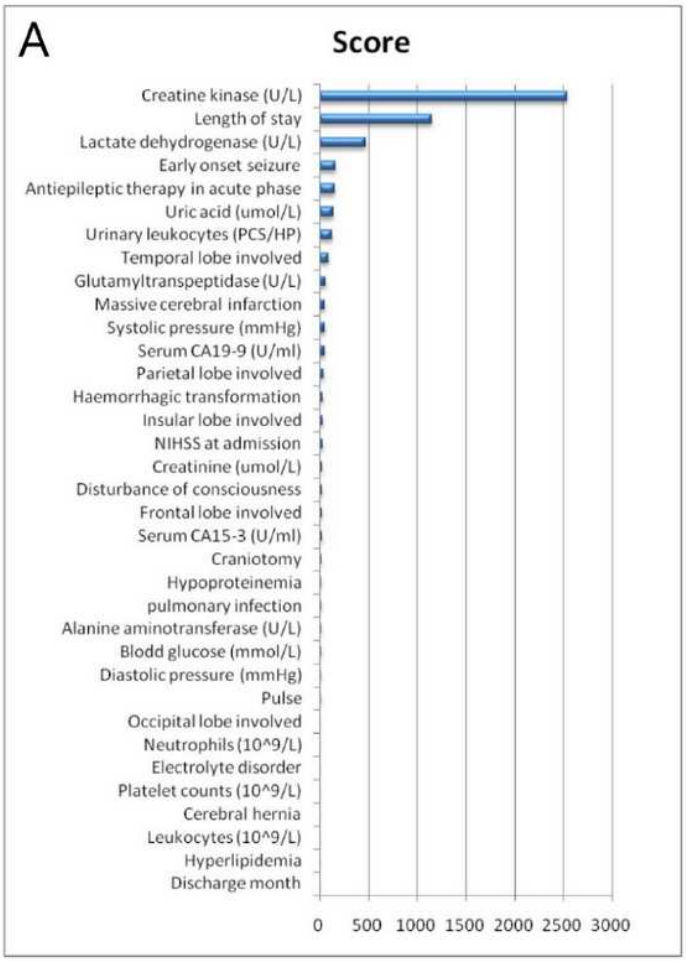


Figure 2

Characteristic variables related to PSE after cerebral infarction (A) and cerebral hemorrhage (B).

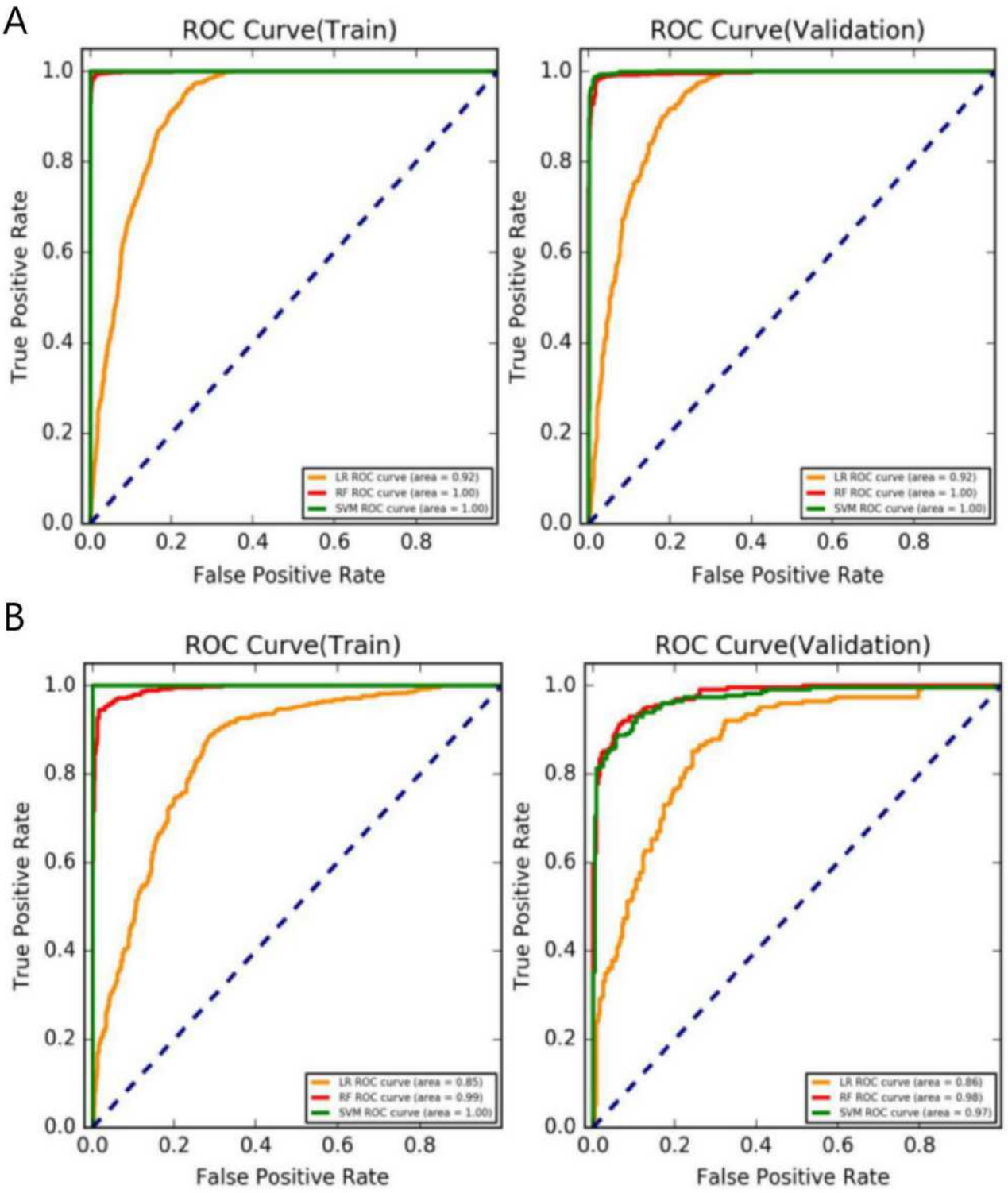


Figure 3

Accuracy of different algorithms in predicting PSE after cerebral infarction (A) and cerebral hemorrhage (B).