Predictive models for secondary epilepsy within 1 year in patients with acute ischemic stroke: a multicenter retrospective study

Jinxin Liu[1,2†], Haoyue He[1,3†],Yanglingxi Wang[1], Jun Du[7],Kaixin Liang[8],Jun Xue[9], Yidan Liang[1], Peng Chen[1], Shanshan Tian[6], Yongbing Deng[1,4,5,]

1 Department of Neurosurgery, Chongqing Emergency Medical Center, Chongqing University Central Hospital, Chongqing, China

2 School of Medicine, Chongqing University, Chongqing, China

3 Bioengineering College of Chongqing University, Chongqing, China

4 Chongqing Key Laboratory of Emergency Medicine

5 Jinfeng Laboratory, Chongqing, China

6 Department of Prehospital Emergency, Chongqing University Central Hospital, Chongqing Emergency Medical Center, Chongqing, China

7 Department of Neurosurgery, Chongqing University Qianjiang Hospital, Chongqing, China

8 Department of Neurosurgery, Yubei District Hospital of Traditional Chinese Medicine, Chongqing, China

9 Department of Neurosurgery, Bishan hospital of Chongqing Medical University, Chongqing, China

8 Chongqing Key Laboratory of Emergency Medicine

[†]These authors have contributed equally to this work and share first authorship.

Corresponding author: Yongbing Deng Email：dyb0913@cqu.edu.cn

Shanshan Tian Email：710836163@qq.com

Author Contributions

JL and HH are both first writer who analysed the data by python and wrote the first draft of the manuscript.YD and TS are both corresponding author who wrote part of the draft and designed the original research. Chongqing University Central Hospital,Chongqing University Qianjiang Hospital , Yubei District hospital and Bishan hospital of Chongqing Medical University provided the database of all cases of the patients. The others collected data and wrote sections of the manuscript. All authors took part in the research and contributed to manuscript revision, read,and approved the submitted version.

Data availability statement

The codes,models,analysis results can be provided for researchers if needed by the corresponding author.

Acknowledgements

The authors would like to thank the colleagues in the information and imaging departments for their hard work contributing to the final research results.

Ethics approval statement

We confirm that we have read the Journal’s position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Funding statement

The research is funded by Based on artificial intelligence and multiple omics technology set up a system of auxiliary cardiovascular disease diagnosis and treatment(2023CDJYGRH-ZD06);by Emergency medicine Key laboratory of Chongqing Joint Fund for Talent Innovation and development （2024RCCX10）,by Brain-like intelligence research Key laboratory of chongqing Education Commission(BIR2019004)

Conflict of interests

The authors have no relevant conflicts of interest to disclose.

Patient consent statement

This study was a retrospective study and only deidentified patient data were collected, exempting the need for patient informed consent rights.

Permission to reproduce material from other sources

There are no reproduce material from other sources.

Clinical trial registration

The trail number is RS202406.

Abstract

Objective:

Post-stroke epilepsy (PSE) is a significant complication that has a negative impact on the prognosis and quality of life of ischemic stroke patients. We collected medical records from 4 hospitals in Chongqing and created an interpretable machine learning model for prediction.

Methods:

We collected medical records, imaging reports, and laboratory tests from 21459 patients with a diagnosis of ischemic stroke . We conducted traditional univariable and multivariable statistics analyses to compare and identify important features. Then the data was divided into a 70% training set and a 30% testing set. We employed the Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors method to resample an imbalanced dataset in the training set. Nine commonly used methods were used to build machine learning models, and relevant prediction metrics were compared to select the best-performing model. Finally, we used SHAP(SHapley Additive exPlanations) for model interpretability analysis, assessing the contribution and clinical significance of different features to the prediction.

Results:

In the traditional regression analysis, complications such as hydrocephalus, cerebral hernia, uremia, deep vein thrombosis; significant brain regions included the involvement of the cortical regions including frontal lobe, parietal lobe, occipital lobe, temporal lobe, subcortical region of basal ganglia, thalamus and so on contributed to PSE. General features such as age, gender, and the National Institutes of Health Stroke Scale score, as well as laboratory indicators including WBC count, D-dimer, lactate, HbA1c and so on were associated with a higher likelihood of PSE. Patients with conditions such as fatty liver, coronary heart disease, hyperlipidemia, and low HDL had a higher likelihood of developing PSE. The machine learning models, particularly tree models such as Random Forest, XGBoost, and LightGBM, demonstrated good predictive performance with an AUC of 0.99.

Conclusion:

The model built on a large dataset can effectively predict the likelihood of PSE, with tree-based models performing the best. The NIHSS score , WBC count and D-dimer were found to have the greatest impact.

Introduction

Stroke is the second leading cause of death worldwide, with an annual mortality rate of approximately 5.5 million, and also the leading cause of disability globally, accounting

for 50% of cases [1]. Generally, ischemic stroke accounts for the majority, about 80% of stroke cases [2][3]. Post-stroke epilepsy (PSE) is a significant complication, with studies indicating that as many as 3-30% of stroke patients develop epilepsy, which has a negative impact on patients' prognosis and quality of life [4]. It can exacerbate cognitive, psychiatric, and physical impairments caused by cerebrovascular disease and comorbidities [5]. Furthermore, the highest incidence of PSE occurs within the first year after acute stroke, accounting for about half of the cases [2]. Therefore, early prediction and intervention for PSE, especially ischemic ones, are crucial.

Currently, most studies utilize clinical data to establish statistical models, survival analysis and cox regression [2][6], and multiple linear regression [7] to construct simple models for the prediction of PSE . Last year, Lin et al. developed a model based on radiomics that outperformed the conventional clinical model in predicting PSE related to intracerebral hemorrhage (ICH). They suggested that a combined radiomics-clinical model could better assist clinicians in assessing the individual risk of PSE after the first occurrence of ICH and facilitate early diagnosis and treatment of PSE [8]. However, subsequent studies have raised doubts regarding the application of radiomics, suggesting the need for further research [9]. Overall, there is still a relative scarcity of research on PSE prediction , with most studies focusing on the analysis of specific or certain risk factors [10][11][8][12] constructing simple models and hardly proposed or established a more comprehensive and scientifically accurate prediction model.

Machine learning has emerged as a promising approach in recent years for constructing medical models, as it excels in handling large volumes of data and complex information, and has been increasingly applied in neuroscience and clinical prediction [13][14][15]. Previous studies have utilized machine learning for related research on post-stroke cognitive impairments [16], stroke and myocardial infarction risk prediction models in large artery vasculitis patients [14], post-stroke depression prediction models based on liver function test indicators [17], and prediction of hematoma expansion in traumatic brain injury (TBI) [18]. Models constructed using machine learning algorithms can automatically handle linear or complex nonlinear relationships between different variables and provide insights into the contribution of different features to the prediction target, which is challenging for traditional statistical models. However, machine learning methods require a substantial amount of data and are prone to overfitting when trained on small sample data. The more valid and high-quality data input, the better machine learning algorithms can capture the underlying patterns between the data, thereby achieving more accurate predictions.

This study try to select important risk factors from mutiple fearures extracted from the clinical records and examination data of ischemic stroke patients and subsequently develops a prediction model for PSE using machine learning methods. By utilizing relevant early admission features of ischemic stroke patients, we aim to automatically predict the probability of PSE occurrence and further guide clinical decision-making and nursing care.

Research content and method

Research patients

This study retrospectively included all stroke patients admitted to the Chongqing Emergency Center between June 2017 and June 2022 for the development of the prediction model. Subsequently, patient data from three external validation centers, namely, Qianjiang Central Hospital, Bishan District People's Hospital, and Yubei District Traditional Chinese Medicine Hospital, were collected between July 2022 and July 2023 for external validation and evaluation of the model. The external validation cohort focused more on collecting positive cases to examine the model's ability to identify positive samples.

Inclusion criteria: (1) Age between 18 and 90 years at admission; (2) Diagnosed with acute ischemic stroke and hospitalized for treatment.

Exclusion criteria: (1) Patients with a history of stroke or transient ischemic attack (TIA); (2) Patients with a history of other conditions such as traumatic brain injury, intracranial tumors, or cerebral vascular malformations that may cause epilepsy; (3) Patients with a history of epilepsy or who have received antiseizure medications for the prevention of seizures or for other diseases (such as migraine or psychiatric disorders); (4) Patients who died within 72 hours after stroke onset.

This study collected de-identified data from relevant patients for the construction of a multi-modal database for stroke patients. The study protocol was approved by the Ethics Committees of Chongqing University Center Hospital, Chongqing University Qianjiang Central Hospital, Bishan District People's Hospital, and Yubei Traditional Chinese Medicine Hospital.

The procedure of selection is in figure1. Total there are 42079 records from the stroke database, 24733 patients were diagonosed as ischemic stroke or lacular stoke with new onset. Then we excluded hemorrage stroke(4565),history of stroke(2154), TIA(3570), unclear cause stroke(561) and records who missed important data(6496). Then we excluded patients whose seizure might be attributed to other potential causes (brain tumor, intracranial vascular malformation, traumatic brain injury,etc)(865). Then we exclude patient who had a seizure history(152) or died in hospital (1444). Then we excluded patients who were lost to follow-up (had no outpatient records and cant contact by phone)or died within 3 months of the stroke incident(813). Finally 21459 cases are involved in this research.

Data collection

We extracted all records and other relevant data from the database of the hospitals. Under the structure of PostgreSQL we coded Structured Query Language to manage different data as follows:

(1) General information: gender, age, NIHSS(the National Institutes of Health Stroke Scale) score at admission;

(2) (2) Comorbidities and complications: uremia, DVT(previous deep vein thrombosis), diabetes mellitus, hypertension, coronary atherosclerosis, atrial fibrillation, cerebral hernia, hydrocephalus, hypoproteinemia, hyperuricemia, hyperlipidemia, internal carotid stenosis, common carotid stenosis,etc.

(3) According to CT or MRI records, the patient's cortical lobes and subcortical involvement were counted: frontal lobe \ parietal lobe \ temporal lobe \ occipital lobe \ insular lobe \ basal ganglia \ internal capsule \ brain stem \ cerebellum \ periventricular \ centrum semiovale \ thalamus involvement. In addition, the extent of cortical involvement (frontal lobe, parietal lobe, temporal lobe, occipital lobe and insular lobe each accumulated 1 point) and the extent of subcortical involvement (basal ganglia, internal capsule, brain stem, periventricular, thalamus and cerebellum any accumulated 1 point) were summarized.

(4) According to CTA, MRA or DSA records, the patient's vascular stenosis or occlusion was counted: ACA(anterior cerebral artery) \ MCA(middle cerebral artery) \ PCA(posterior cerebral artery) \ VA(vertebral artery) \ BA(basilar artery)

(5) Important laboratory indicators: Blood lipids (TG(Triglyceride), HDL(High Density Lipoprotein Cholesterol), LDL(Low Density Lipoprotein Cholesterol)), liver function (ALT(Alanine Transaminase), AST(Aspartate Aminotransferase), Bilirubin, Albumin), renal function (Urea, BUA(Blood Uric Acid), Creatinine), blood gas (Lactate, Anion Gap, TCO2(Total Carbon Dioxide)), coagulation related indicators (INR(International Normalized Ratio), PT(Prothrombin Time), APTT(Activated Partial Thromboplastin Time), TT(Thrombin Time), D-Dimer, Fibrinogen) and myocardial enzymes (CK(Creatine Kinase), CK-MB(Creatine Kinase Isoenzyme), LDH(Lactate Dehydrogenase), IMA(Ischemic Modified Albumin), HBDH(α-Hydroxybutyrate Dehydrogenase)).

Data processing and model building

(Processing of missing data) We counted the values of all laboratory indicators for the first time after stroke admission(everyone who was admitted because of stroke would perform blood routine , liver and kidney function and so on), excluded indicators with missing values of more than 10%, and filled the data of the remaining indicators with missing values by random forest algorithm using the default parameter. First, we go through all the features, starting with the one with the least missing (since the least accurate information is needed to fill in the feature with the least missing). When filling in a feature, replace the missing value of the other feature with 0. Each time a regression prediction is completed, the predicted value is placed in the original feature matrix and the next feature is filled in. After going through all the features, the data is complete.

(Distribution of characteristics) Univariate analysis was used to examine the distribution of characteristics between the PSE negative group and the positive group. The data were then divided into a training set and a test set by .

(Processing of unbalanced data) Considering the low incidence of PSE and the small proportion of positive patients, the positive data of the training set were augmented by Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors by using default parameter of SMOTEENN method from imblearn python package and set random seed at 42 for repetition.

(Processing of categorical data) For categorical data, the one-hot method is used for transformation. The LASSO method was then used in the training set to screen the important features.

(Model building) We first used LASSO regression to select the 20 most important features. Next, we employed 9 common machine learning methods, including Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Multi-Layer Perceptron, XGBoost, LightGBM, and K-Nearest Neighbors. We then optimized the hyperparameters of each model through grid search to improve their performance.To evaluate the models, we calculated metrics such as accuracy, sensitivity, specificity, F1-score, positive predictive value, and negative predictive value. We also plotted the ROC curve, calibration curve, and decision curve. Additionally, we used an independent external validation dataset to assess the generalization performance of the selected model. Finally, we leveraged the SHAP algorithm to perform an interpretable analysis of the best-performing model, investigating the contribution of each feature to the model's predictions and their clinical significance. Through this series of model building, optimization, and analysis, we developed a machine learning model with good predictive performance and interpretability, providing valuable support for clinical decision-making.

Statistical approach

PostgreSQL v15 (http://www.postgresql.org/) was used to search and extract the data from the local database.

The open-source statistical package "Scipy.stats" in Python was used for statistical analysis. The details of the univariate significance analysis for each feature are as follows:

First, the Shapiro-Wilk test was used to check the normality of the distribution for each feature. For features that did not follow a normal distribution, the Mann-Whitney U test was used to assess their significance with respect to the target variable.

For features that exhibited a normal distribution, the Levene test was employed to assess the homogeneity of variances. Features with homogeneous variances were analyzed using the Student's t-test to determine their significance with respect to the target variable, while features with heterogeneous variances were analyzed using the Welch's t-test.

The confidence intervals for the AUC values and Brier scores were obtained by performing 1000 bootstrap resampling iterations on the corresponding datasets. The binary classification thresholds for the predicted probabilities generated by all models were established using the maximum Youden index derived from the training cohort.

Throughout the study, a two-tailed p-value less than 0.05 was considered statistically significant.

All the codes were uploaded at https://github.com/conanan/lasso-ml.

Results

Filling of missing data

These features had missing values that were filled using a Random Forest (RF) model, addressing the missing data one feature at a time: Plt, WBC, RBC, HbA1c, CRP, TG, LDL, HDL, AST, ALT, Bilirubin, Albumin, Urea, Creatinine, BUA, PT, APTT, TT, INR, D-dimer, Fibrinogen, CK, CK-MB, LDH, HBDH, IMA, Lactate, Anion_gap, TCO2, NIHSS.

Characteristics of study participants

A total of 21459 patients were included in this study, of which 15021 patients were included in the training set, and the incidence of PSE was 4.3%. The test set contained 6438 patients with a PSE incidence of 4.3%. The external validation cohort consisted of 536 patients at three hospitals. Statistical details of the clinical characteristics of the patients are provided in the table1.

Statistical analysis showed that the patients who had higher possibility of PSE were with complications of uremia, history of DVT, atrial fibrillation, hyperuricemia, cerebral hernia and hydrocephalus. The involved locations of frontal lobe, parietal lobe, occipital lobe, temporal lobe, cortex, subcortex, basal ganglia and hypothalamus. The general characteristics included age, gender, nihss score; Laboratory indicators included wbc count, hba1c, crp, tg, ast, alt, bilirubin, urea, bua, aptt, tt, d_dimer, ck, ckmb, ldh, hbdh, ima, lactate, and anion_gap . Besides, the p values of fatty liver, coronary heart disease, hyperlipidemia, and hdl were significant, and patients with negative or low values of these indicators had a high risk of secondary disease. The statistics analysis result, the uni and multi regression analysis result table is in table1,table 2 and table 3.

Performance of machine learning models

The relevant indicators of the machine learning model are shown in table4, and the ROC curves, calibration curve and DCA are shown in figure3. It can be found that the over all models the AUC of tree models such as RF, XGboost and lightGBM are better than other models, and the PPV value of random forest is the highest, reaching 0.864, which is the most important function of our models. Complex machine learning algorithms were superior to traditional logistic regression. The Brier score of the calibration curve reached 0.006, and the DCA also showed good clinical decision-making benefits, which had good practical value. In the external validation cohort, we use the RF to predict. The Sensitivity was 0.91, the PPV was 0.95, demonstrate a good predictive ability of the model.

Analysis of SHAP risk factors

The analysis in Figure 4 shows the SHAP (Shapley Additive Explanations) values, individual decision attempts, and overall decision curves. Among the general characteristics, females had a higher rate of PSE.

Regarding the NIHSS score, higher score cause higher incidence rate of PSE .Higher values of WBC count, D-dimer, CRP , AST , CK-MB, HbA1c, bilirubin, TCO2, and LDH at admission were associated with a greater likelihood of developing PSE. Conversely, lower values of HBDH , PLT, and APTT were also linked to a higher probability of the outcome.However, the specific regions of the brain affected did not have a significant individual effect on the overall outcome.Among the complications, only hypertension was more strongly associated with the development of the outcome. Other conditions, such as coronary heart disease, diabetes, hyperlipidemia, and fatty liver, were less likely to be related to the outcome. We use the force plot of the first person to show the influence of different features of the first person, we can see that long APTT time contribute best to PSE, then the AST level and others, the NIHSS score may be low and contribute opposite to the final result. Then the decision plot is a collection of model decisions that show how complex models arrive at their predictions.

Discussion

Our study utilized comprehensive clinical data, imaging data, laboratory test data, from the database of the stroke patients and employed machine learning algorithms to establish a predictive model, achieving an AUC score of above 0.95, which demonstrated more accurate predictions compared to traditional statistical methods. Our research found that tree-based ensemble models showed superior overall prediction capabilities when dealing with large sample sizes and high-dimensional features.

During the modeling process, due to the extreme imbalance between negative and positive samples, we employed SMOTEENN technique to resample an imbalanced dataset for machine learning, resulting in improved training performance. Through SHAP analysis, we conducted interpretability analysis of the model and determined the importance of different features.

In our study, age and NIHSS score were treated as continuous variables. We found that, overall, female patients, older patients, and those with higher NIHSS scores were more prone to develop PSE, which is consistent with recent articles. High NIHSS scores, indicative of more severe stroke, increased the risk of complications, ranking only to white blood cell count and d-dimer in our model [5][19][10][20]. However, there are conflicting opinions regarding the impact of age. Some researchs [5][21] suggested that age <65 is a high-risk factor, which aligns with our findings, while some studies [22] confirmed that advanced age is the determining factor. Yamada et al. [21] also concurred with our study in identifying a higher risk of complications among females, whereas Waafi et al. [10] indicated that the likelihood of male patients developing complications is 3.325 times that of females, which contradicts our findings.

Previous studies have shown that patients with diabetes, dyslipidemia, hypertension, depression, or dementia are at an increased risk of developing vascular epilepsy [12]. In our study, statistics and multiple ML models analyzed the association between comorbidities and complications, revealing that patients with coronary heart disease, diabetes, fatty liver, hyperlipidemia, or large artery stenosis or plaques(CCA and ICA) were less likely to develop epilepsy. According to the TOAST classification, ischemic stroke is categorized into five types: large artery atherosclerosis, cardioembolism, small vessel occlusion, other determined etiology, and undetermined etiology. Patients with combined comorbidities generally fall into the categories of large artery atherosclerosis and cardioembolism, which are relatively well-defined and easier to intervene, thus resulting in a lower likelihood of developing epilepsy. Conversely, strokes with undetermined etiology usually have a poor prognosis and are more likely to lead to epilepsy. Among diabetes patients, higher HbA1c levels indicate poorer blood sugar control, resulting in a higher probability of developing complications, which significantly affects certain patients, while those with good control have a lower overall risk of developing complications.

Alain et al. found that cortical infarction was more likely to result in epilepsy in patients hospitalized with anterior circulation ischemic stroke [23]. Lin et al. found that factors such as cortical involvement and intracerebral hemorrhage volume increased the likelihood of PSE, which is consistent with our research findings [8]. Al-Sahli et al. also suggested that cortical brain injury and large-area lesions increased the risk of PSE [5][21]. In our study, statistics showed affections of cortical and subcortical regions both increased the possibility of PSE, but had lower affection than the other features so didn't be selected in lasso regression.

Previous studies have found that acute infection is a risk factor for ischemic stroke [24]. C-reactive protein (CRP) reflects the level of inflammation and is an independent prognostic factor [25]. In our study, regression and SHAP analysis both showed that white blood cell count had great impact among the routine blood test parameters, in SHAP it even surpassed the NIHSS score. High white blood cell count may indicate severe inflammation and infection, as well as increased blood viscosity, making patients more susceptible to secondary complications. In general, high red blood cell count and low platelet count also have some influence.

A large-scale study on Chinese individuals found a negative correlation between plasma high-density lipoprotein cholesterol (HDL-C) concentration and the risk of ischemic stroke, a weak positive correlation between plasma triglyceride (TG) concentration and the risk of ischemic stroke, and a strong correlation between plasma low-density lipoprotein cholesterol (LDL-C) concentration and apolipoprotein B [26]. High HDL-C levels are associated with better prognosis [27]. Our study is consistent with previous research, indicating that high LDL-C, low HDL-C, and high TG levels are more likely to lead to PSE. This can be easily understood as high cholesterol and triglyceride levels lead to increased blood viscosity and vascular sclerosis, making it easier for clots to form [12][28][29]. Higher D-dimer levels indicate greater brain tissue damage and a higher likelihood of PSE. Overall, lower activated partial thromboplastin time (APTT) and fibrinogen levels are associated with an increased risk of PSE. INR, PT, and TT have a smaller impact. Among liver function parameters, aspartate aminotransferase (AST) has the greatest influence on PSE, while high AST levels, low alanine aminotransferase (ALT) levels, and low albumin levels all have a certain degree of impact. Lingling Ding et al. found that liver enzyme subgroups characterized by alanine aminotransferase and aspartate aminotransferase were associated with a high risk of adverse function [30], which is consistent with our research.

Studies have shown that subgroups identified by renal function biomarkers such as urinary microalbumin, cystatin C, and creatinine have significantly higher stroke recurrence and poorer prognosis [30]. In our study, low urea levels and high uric acid levels had a negative impact [31][32][33]. Our research is similar to their conclusions. While elevated uric acid levels at admission are positively associated with PSE, patients previously diagnosed with hyperuricemia are less likely to develop epilepsy. Considering that uric acid functions as a strong reducing agent and has neuroprotective properties [34], patients with normal liver and kidney function and a certain degree of hyperuricemia have stronger resistance to emergencies [35][36]. However, excessively high uric acid levels indicate metabolic disorders and poor liver and kidney function, which are associated with poor prognosis.

When stroke patients are admitted, cardiac enzyme profile tests are often performed to rule out concurrent myocardial ischemia. However, studies have shown that elevated CK-MB in stroke patients may not only be related to the heart [37]. Multiple cardiac enzymes are important prognostic indicators [38][39] and have been included in stroke scores [40]. Some studies have shown a higher incidence of abnormal serum cardiac enzyme profiles in the acute phase of stroke. Although the incidence of abnormalities is unrelated to the nature of the stroke, it is associated with the severity of the stroke, with patients with consciousness disorders having a significantly higher incidence of abnormal cardiac enzyme profiles than those without consciousness disorders [41]. In our study, CK, CK-MB, and IMA in the cardiac enzyme profile had a significant impact and high predictive value, but the specific mechanisms require further research [34].

Although our study incorporates a large amount of information and utilizes almost all available data, including clinical data, imaging data, and laboratory test data, in an attempt to establish more accurate prediction models beyond traditional statistics using machine learning algorithms, there are still several limitations in the modeling process.

While the current study provides valuable insights, the data sample may not be fully representative, and the model's generalization ability requires further assessment. Although the data was collected from multiple tertiary hospitals, encompassing over 20,000 cases, earlier data was lost due to hospital system upgrades. The collected data mainly represents patients diagnosed in the past five years and is primarily concentrated in the Chongqing region, which may limit the model's applicability to other geographic areas.

Additionally, the retrospective nature of the research has resulted in the lack of certain important predictive indicators. As this was a retrospective study, many potentially meaningful features, such as hemorheology, thromboelastography, and hormone levels, were significantly missing and had to be excluded. Incorporating these additional features could potentially improve the model's accuracy.

To enhance the predictive power of the model, it would be beneficial to incorporate more beyond baseline patient characteristics. For example, the current analysis primarily utilized the results of the first examination upon admission, without fully leveraging the information from subsequent examinations. In future research, the use of recurrent neural networks could facilitate the comprehensive extraction of features from the entire sequence of examinations.

To further strengthen the study, data standardization should be improved, and the number of cases and important indicators should continue to increase. Additionally, it would be advisable to explore more advanced scientific methods, such as deep learning, and fully leverage all available data to make more accurate predictions.

Conclusion

We developed an interpretable machine learning model to predict the risk of post-stroke epilepsy (PSE) in hospitalized patients with ischemic stroke. Leveraging a large volume of medical records, our artificial intelligence model demonstrates good predictive performance for PSE.The key predictors identified by the model include NIHSS score, D-dimer levels, lactate levels, and white blood cell count, followed by indicators related to liver function and cardiac enzyme profiles. The transparency and interpretability of the model's predictions can foster trust among clinical practitioners and facilitate decision-making.While the results are promising, further prospective studies are needed to validate the clinical utility of this tool before its application in real-world settings.

[1] Feigin V L, Krishnamurthi R V, Theadom A M, et al.. Global, Regional, and National Burden of Neurological Disorders during 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015[J]. The Lancet Neurology, 2017, 16(11): 877–897.

[2] Galovic M, Döhler N, Erdélyi-Canavese B, 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[J]. The Lancet Neurology, 2018, 17(2): 143.

[3] Krishnamurthi R V, Feigin V L, Forouzanfar M H, et al.. Global and Regional Burden of First-Ever Ischaemic and Haemorrhagic Stroke during 1990–2010: Findings from the Global Burden of Disease Study 2010[J]. The Lancet Global Health, 2013, 1(5): e259–e281.

[4] Zhao Y, Li X, Zhang K, et al.. The Progress of Epilepsy after Stroke[J]. Curr Neuropharmacol, 2018, 16(1): 71–78.

[5] Al-Sahli O a M, Tibekina L, Subbotina O P, et al.. Post-Stroke Epileptic Seizures: Risk Factors, Clinical Presentation, Principles of Diagnosis and Treatment[J]. Epilepsy and paroxysmal conditions, 2023, 15(2): 148–159.

[6] Chen Z, Churilov L, Chen Z, et al.. Association between Implementation of a Code Stroke System and Poststroke Epilepsy[J]. Neurology, 2018, 90(13): e1126–e1133.

[7] Merkler A E, Gialdini G, Lerario M P, et al.. Population-Based Assessment of the Long-Term Risk of Seizures in Survivors of Stroke[J]. Stroke, 2018, 49(6): 1319–1324.

[8] Lin R, Lin J, Xu Y, et al.. Development and Validation of a Novel Radiomics-Clinical Model for Predicting PSE after First-Ever Intracerebral Haemorrhage[J]. European Radiology, 2023, 33(7): 4526–4536.

[9] Pszczolkowski S, Law Z K. Editorial Comment on 《Development and Validation of a Novel Radiomics-Clinical Model for Predicting PSE after First-Ever Intracerebral Haemorrhage》[J]. European Radiology, 2023, 33(7): 4524–4525.

[10] Waafi A K, Husna M, Damayanti R, et al.. Clinical Risk Factors Related to PSE Patients in Indonesia: A Hospital-Based Study[J]. Egyptian Journal of Neurology, Psychiatry and Neurosurgery, 2023, 59(1).

[11] Herzig-Nichtweiß J, Salih F, Berning S, et al.. Prognosis and Management of Acute Symptomatic Seizures: A Prospective, Multicenter, Observational Study[J]. Annals of Intensive Care, 2023, 13(1).

[12] Pitkänen A, Roivainen R, Lukasiuk K. Development of Epilepsy after Ischaemic Stroke[J]. The Lancet Neurology, 2016, 15(2): 185–197.

[13] The Artificial Intelligence Revolution in Stroke Care: A Decade of Scientific Evidence in Review[J]. World Neurosurgery, Elsevier, 2024.

[14] Predicting Stroke and Myocardial Infarction Risk in Takayasu Arteritis with Automated Machine Learning Models[J]. iScience, Elsevier, 2023, 26(12): 108421.

[15] Daidone M, Ferrantelli S, Tuttolomondo A, et al.. Machine Learning Applications in Stroke Medicine: Advancements, Challenges, and Future Prospective[J]. Neural Regeneration Research, 2024, 19(4): 769–773.

[16] Lee M, Yeo N-Y, Ahn H-J, et al.. Prediction of Post-Stroke Cognitive Impairment after Acute Ischemic Stroke Using Machine Learning[J]. Alzheimer's Research and Therapy, 2023, 15(1).

[17] Gong J, Zhang Y, Zhong X, et al.. Liver Function Test Indices-Based Prediction Model for Post-Stroke Depression: A Multicenter, Retrospective Study[J]. BMC Medical Informatics and Decision Making, 2023, 23(1).

[18] He H, Liu J, Li C, et al.. Predicting Hematoma Expansion and Prognosis in Cerebral Contusions: A Radiomics-Clinical Approach[J]. Journal of Neurotrauma, 2024: neu.2023.0410.

[19] Lin R, Yu Y, Wang Y, et al.. Risk of PSE Following Stroke-Associated Acute Symptomatic Seizures[J]. Frontiers in Aging Neuroscience, 2021, 13.

[20] Zöllner J P, Misselwitz B, Kaps M, et al.. National Institutes of Health Stroke Scale (NIHSS) on Admission Predicts Acute Symptomatic Seizure Risk in Ischemic Stroke: A Population-Based Study Involving 135,117 Cases[J]. Scientific Reports, 2020, 10(1).

[21] Yamada S, Nakagawa I, Tamura K, et al.. Investigation of Poststroke Epilepsy (INPOSE) Study: A Multicenter Prospective Study for Prediction of Poststroke Epilepsy[J]. J Neurol, 2020, 267(11): 3274–3281.

[22] Lidetu T, Zewdu D. Incidence and Predictors of Post Stroke Seizure among Adult Stroke Patients Admitted at Felege Hiwot Compressive Specialized Hospital, Bahir Dar, North West Ethiopia, 2021: A Retrospective Follow up Study[J]. BMC Neurology, 2023, 23(1).

[23] Lekoubou A, Ssentongo P, Maffie J, et al.. Associations of Small Vessel Disease and Acute Symptomatic Seizures in Ischemic Stroke Patients[J]. Epilepsy & Behavior, 2023, 145: 109233.

[24] Bova I Y, Bornstein N M, Korczyn. Acute Infection as a Risk Factor for Ischemic Stroke[J]. Stroke, 1996, 27(12): 2204–2206.

[25] Di Napoli M, Papa F, Bocola V. C-Reactive Protein in Ischemic Stroke an Independent Prognostic Factor[J]. Stroke, 2001, 32(4): 917–924.

[26] Sun L, Clarke R, Bennett D, et al.. Causal Associations of Blood Lipids with Risk of Ischemic Stroke and Intracerebral Hemorrhage in Chinese Adults[J]. Nat Med, Nature Publishing Group, 2019, 25(4): 569–574.

[27] Bandeali S, Farmer J. High-Density Lipoprotein and Atherosclerosis: The Role of Antioxidant Activity[J]. Current Atherosclerosis Reports, 2012, 14(2): 101–107.

[28] Gasparini S, Neri S, Brigo F, et al.. Late Epileptic Seizures Following Cerebral Venous Thrombosis: A Systematic Review and Meta-Analysis[J]. Neurol Sci, 2022, 43(9): 5229–5236.

[29] Abraira L, Giannini N, Santamarina E, et al.. Correlation of Blood Biomarkers with Early-Onset Seizures after an Acute Stroke Event[J]. Epilepsy & Behavior, 2020, 104: 106549.

[30] Ding L, Liu Y, Meng X, et al.. Biomarker and Genomic Analyses Reveal Molecular Signatures of Non-Cardioembolic Ischemic Stroke[J]. Sig Transduct Target Ther, Nature Publishing Group, 2023, 8(1): 1–16.

[31] Zhang W, Cheng Z, Fu F, et al.. Serum Uric Acid and Prognosis in Acute Ischemic Stroke: A Dose–Response Meta-Analysis of Cohort Studies[J]. Frontiers in Aging Neuroscience, 2023, 15.

[32] Wang D, Hu B, Dai Y, et al.. Serum Uric Acid Is Highly Associated with Epilepsy Secondary to Cerebral Infarction[J]. Neurotox Res, 2019, 35(1): 63–70.

[33] Wang C, Cui T, Wang L, et al.. Prognostic Significance of Uric Acid Change in Acute Ischemic Stroke Patients with Reperfusion Therapy[J]. Eur J Neurol, 2021, 28(4): 1218–1224.

[34] Ng G J L, Quek A M L, Cheung C, et al.. Stroke Biomarkers in Clinical Practice: A Critical Appraisal[J]. NeuROChemistry International, 2017, 107: 11–22.

[35] Amaro S, Urra X, Gómez-Choco M, et al.. Uric Acid Levels Are Relevant in Patients With Stroke Treated With Thrombolysis[J]. Stroke, American Heart Association, 2011, 42(1_suppl_1): S28–S32.

[36] Amaro S, Urra X, Gómez-Choco M, et al.. Uric Acid Levels Are Relevant in Patients with Stroke Treated with Thrombolysis[J]. Stroke, 2011, 42(SUPPL. 1): S28–S32.

[37] Ay H, Arsava E M, Sarba O. Creatine Kinase-MB Elevation after Stroke Is Not Cardiac in Origin Comparison with Troponin T Levels[J]. Stroke, 2002, 33(1): 286–289.

[38] Liu X, Chen X, Wang H, et al.. Prognostic Significance of Admission Levels of Cardiac Indicators in Patients with Acute Ischaemic Stroke: Prospective Observational Study[J]. J Int Med Res, SAGE Publications Ltd, 2014, 42(6): 1301–1310.

[39] Zeng Y-Y, Zhang W-B, Cheng L, et al.. Cardiac Parameters Affect Prognosis in Patients with Non-Large Atherosclerotic Infarction[J]. Molecular Medicine, 2021, 27(1): 2.

[40] Hijazi Z, Lindbäck J, Alexander J H, et al.. The ABC (Age, Biomarkers, Clinical History) Stroke Risk Score: A Biomarker-Based Risk Score for Predicting Stroke in Atrial Fibrillation[J]. European Heart Journal, 2016, 37(20): 1582–1590.

[41] Zheng Yuan-Hui, ZHENG Jin-Yi, ZHANG Jian. Changes of serum myocardial enzyme profile in acute stage of stroke [J]. Chinese Journal of Advanced Medical Doctors, China Medical Journal, 2009, 32(07): 46 -- 47.

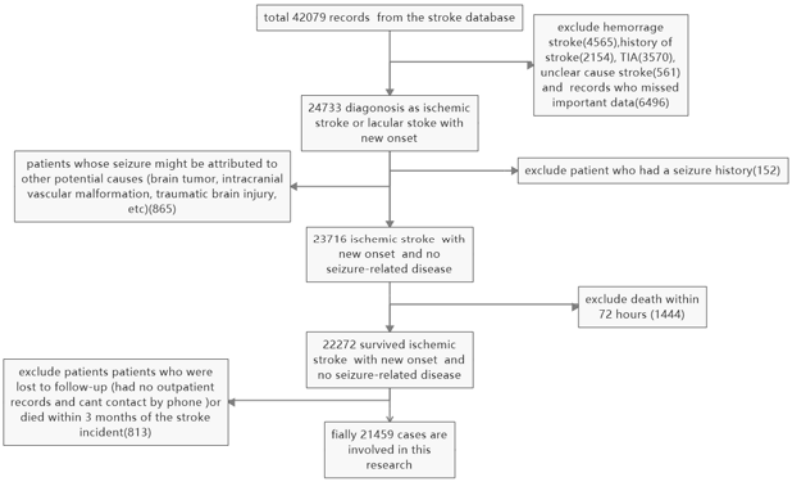

Figure 1.Selection and exclusion procedure of patients

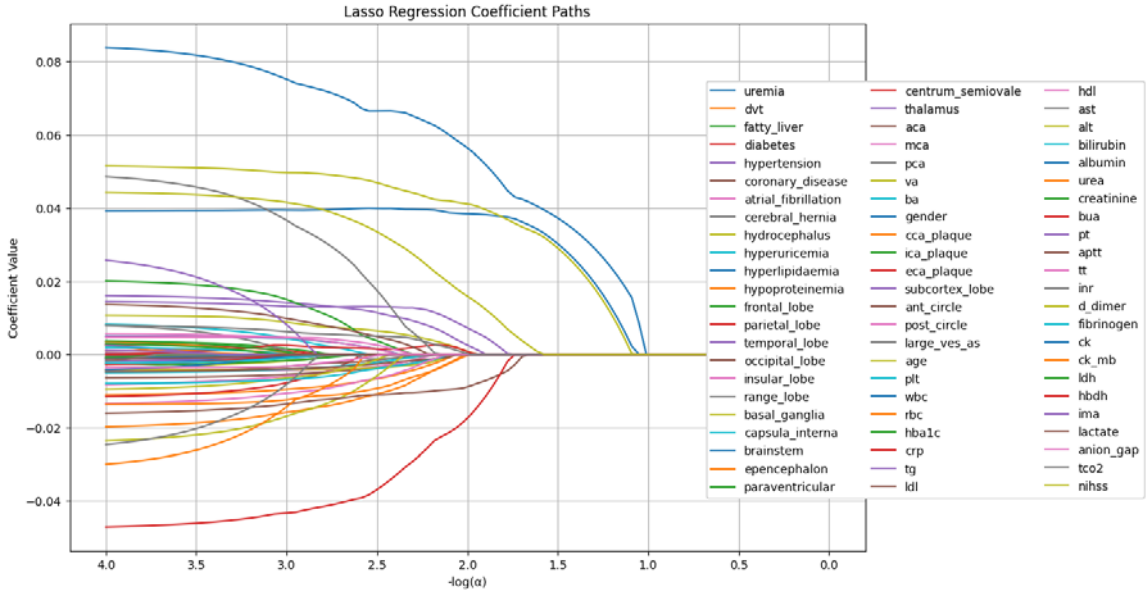

Figure 2.LASSO Regression Coefficient Paths

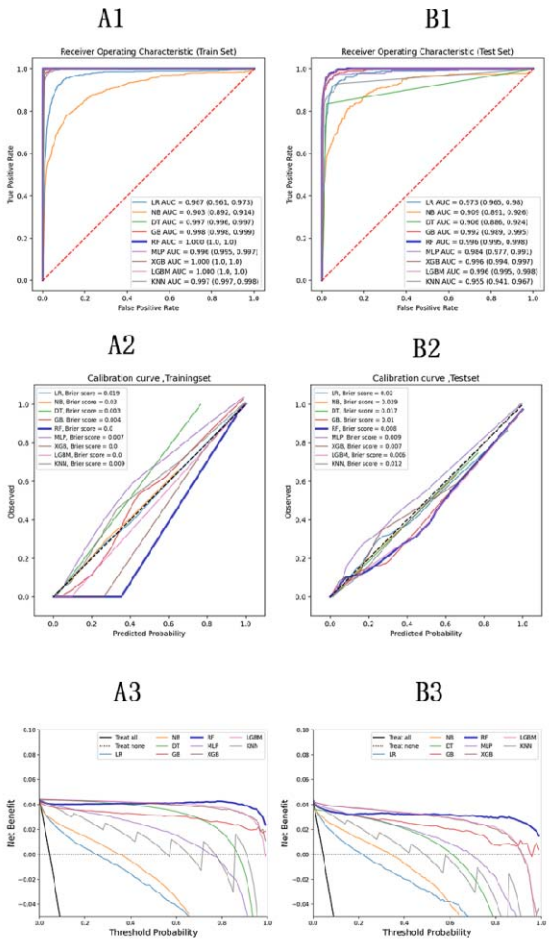

Figure 3. ROC of train(A1) ,ROC of test(B1), CC of train(A2) ,CC of test(B2), PCA of train(A3),PCA of test(B3)

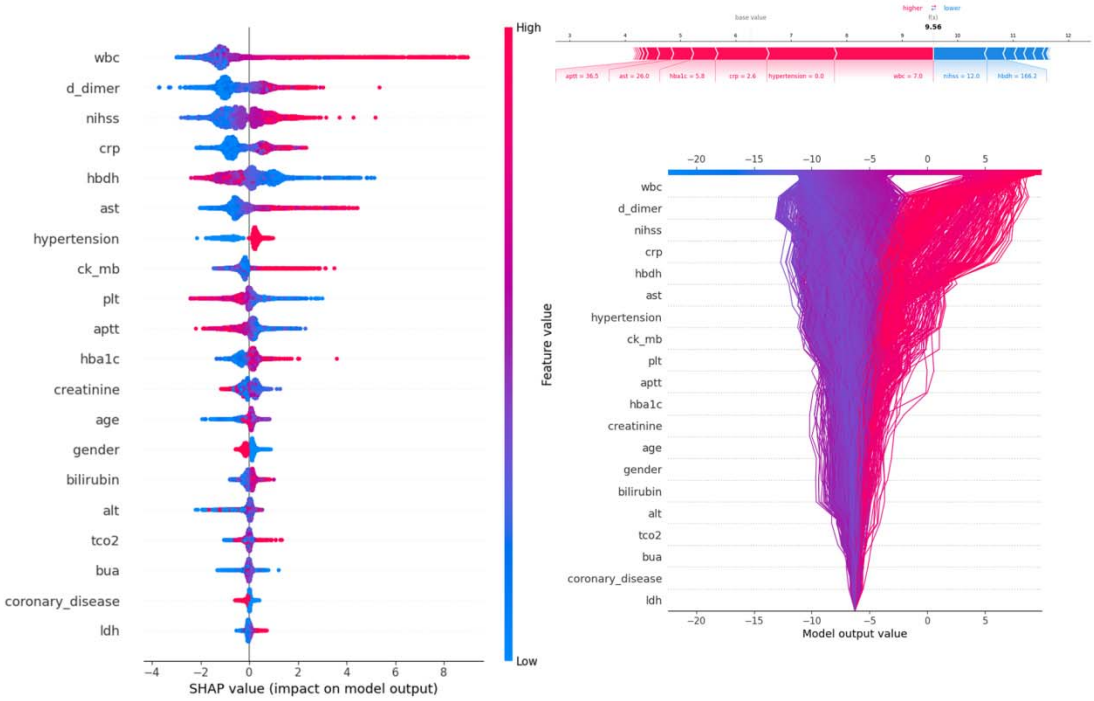

Figure 4.SHAP value(left),force plot(upper right) and decision plot (lower right)

Feature	positive, N=954	negative, N = 20789	method	P	stats
eca_plaque	-	-	Chi-Square	0.438971	0.59897
——0	942 (98.742%)	20591 (99.048%)	-	-	-
——1	12 (1.258%)	198 (0.952%)	-	-	-
subcortex_lobe	-	-	Chi-Square	0.001273	10.381551
——0	814 (85.325%)	18454 (88.768%)	-	-	-
——1	140 (14.675%)	2335 (11.232%)	-	-	-
ba	-	-	Chi-Square	0.991017	0.000127
——0	945 (99.057%)	20605 (99.115%)	-	-	-
——1	9 (0.943%)	184 (0.885%)	-	-	-
hypertension	-	-	Chi-Square	0.602539	0.271184

——0	290 (30.398%)	6497 (31.252%)	-	-	-
——1	664 (69.602%)	14292 (68.748%)	-	-	-
ica_plaque	-	-	Chi-Square	0.152086	2.051203
——0	878 (92.034%)	19392 (93.28%)	-	-	-
——1	76 (7.966%)	1397 (6.72%)	-	-	-
frontal_lobe	-	-	Chi-Square	0	53.171781
——0	868 (90.985%)	19943 (95.931%)	-	-	-
——1	86 (9.015%)	846 (4.069%)	-	-	-
cerebral_hernia	-	-	Chi-Square	0.000032	17.284355
——0	934 (97.904%)	20626 (99.216%)	-	-	-
——1	20 (2.096%)	163 (0.784%)	-	-	-
thalamus	-	-	Chi-Square	0.060918	3.512207
——0	937 (98.218%)	20565 (98.923%)	-	-	-
——1	17 (1.782%)	224 (1.077%)	-	-	-
occipital_lobe	-	-	Chi-Square	0.000034	17.17679
——0	919 (96.331%)	20422 (98.235%)	-	-	-
——1	35 (3.669%)	367 (1.765%)	-	-	-
pca	-	-	Chi-Square	0.891182	0.018717
——0	952 (99.79%)	20729 (99.711%)	-	-	-
——1	2 (0.21%)	60 (0.289%)	-	-	-
paraventricular	-	-	Chi-Square	0.213759	1.545786
——0	899 (94.235%)	19786 (95.175%)	-	-	-
——1	55 (5.765%)	1003 (4.825%)	-	-	-
mca	-	-	Chi-Square	0.393066	0.729435
——0	912 (95.597%)	19998 (96.195%)	-	-	-
——1	42 (4.403%)	791 (3.805%)	-	-	-
coronary_disease	-	-	Chi-Square	0	26.19087

——0	599 (62.788%)	11288 (54.298%)	-	-	-
——1	355 (37.212%)	9501 (45.702%)	-	-	-
hypoproteinemia	-	-	Chi-Square	0	53.351931
——0	774 (81.132%)	18479 (88.888%)	-	-	-
——1	180 (18.868%)	2310 (11.112%)	-	-	-
parietal_lobe	-	-	Chi-Square	0	57.137771
——0	884 (92.662%)	20180 (97.071%)	-	-	-
——1	70 (7.338%)	609 (2.929%)	-	-	-
aca	-	-	Chi-Square	0.928981	0.007944
——0	941 (98.637%)	20524 (98.725%)	-	-	-
——1	13 (1.363%)	265 (1.275%)	-	-	-
brainstem	-	-	Chi-Square	0.294979	1.096759
——0	938 (98.323%)	20532 (98.764%)	-	-	-
——1	16 (1.677%)	257 (1.236%)	-	-	-
hyperuricemia	-	-	Chi-Square	0.000001	25.147468
——0	801 (83.962%)	18547 (89.215%)	-	-	-
——1	153 (16.038%)	2242 (10.785%)	-	-	-
temporal_lobe	-	-	Chi-Square	0	57.872112
——0	886 (92.872%)	20209 (97.21%)	-	-	-
——1	68 (7.128%)	580 (2.79%)	-	-	-
diabetes	-	-	Chi-Square	0.389926	0.739172
——0	617 (64.675%)	13737 (66.078%)	-	-	-
——1	337 (35.325%)	7052 (33.922%)	-	-	-
range_lobe	-	-	Chi-Square	0	85.377485
——0	830 (87.002%)	19559 (94.083%)	-	-	-
——1	43 (4.507%)	467 (2.246%)	-	-	-
——2	32 (3.354%)	329 (1.583%)	-	-	-

——3	31 (3.249%)	224 (1.077%)	-	-	-
——4	15 (1.572%)	175 (0.842%)	-	-	-
——5	3 (0.314%)	35 (0.168%)	-	-	-
epencephalon	-	-	Chi-Square	1	0
——0	934 (97.904%)	20362 (97.946%)	-	-	-
——1	20 (2.096%)	427 (2.054%)	-	-	-
hydrocephalus	-	-	Chi-Square	0	181.23517
——0	895 (93.816%)	20565 (98.923%)	-	-	-
——1	59 (6.184%)	224 (1.077%)	-	-	-
insular_lobe	-	-	Chi-Square	0.391042	0.735699
——0	938 (98.323%)	20519 (98.701%)	-	-	-
——1	16 (1.677%)	270 (1.299%)	-	-	-
gender	-	-	Chi-Square	0	44.244052
——0	372 (38.994%)	10407 (50.06%)	-	-	-
——1	582 (61.006%)	10382 (49.94%)	-	-	-
uremia	-	-	Chi-Square	0.00008	15.568169
——0	934 (97.904%)	20618 (99.177%)	-	-	-
——1	20 (2.096%)	171 (0.823%)	-	-	-
atrial_fibrillation	-	-	Chi-Square	0.008017	7.029734
——0	838 (87.841%)	18811 (90.485%)	-	-	-
——1	116 (12.159%)	1978 (9.515%)	-	-	-
centrum_semiovale	-	-	Chi-Square	0.36206	0.830735
——0	922 (96.646%)	20207 (97.2%)	-	-	-
——1	32 (3.354%)	582 (2.8%)	-	-	-
basal_ganglia	-	-	Chi-Square	0.005355	7.755329
——0	893 (93.606%)	19869 (95.575%)	-	-	-
——1	61 (6.394%)	920 (4.425%)	-	-	-

dvt	-	-	Chi-Square	0	40.790867
——0	847 (88.784%)	19534 (93.963%)	-	-	-
——1	107 (11.216%)	1255 (6.037%)	-	-	-
fatty_liver	-	-	Chi-Square	0.000171	14.123893
——0	812 (85.115%)	16655 (80.114%)	-	-	-
——1	142 (14.885%)	4134 (19.886%)	-	-	-
hyperlipidaemia	-	-	Chi-Square	0.000317	12.969155
——0	801 (83.962%)	16439 (79.075%)	-	-	-
——1	153 (16.038%)	4350 (20.925%)	-	-	-
cca_plaque	-	-	Chi-Square	0.376965	0.780577
——0	751 (78.721%)	16100 (77.445%)	-	-	-
——1	203 (21.279%)	4689 (22.555%)	-	-	-
va	-	-	Chi-Square	0.797483	0.065847
——0	927 (97.17%)	20159 (96.97%)	-	-	-
——1	27 (2.83%)	630 (3.03%)	-	-	-
fibrinogen	3.518 ± 0.663	3.602 ± 0.464	Mann-Whitney U	0.434584	10064078.5
d_dimer	4.362 ± 4.398	1.198 ± 0.98	Mann-Whitney U	0	3555180.5
bua	342.521 ± 74.651	344.132 ± 58.336	Mann-Whitney U	0.000037	10698805.5
tco2	22.739 ± 1.025	22.781 ± 1.225	Mann-Whitney U	0.166751	10178363
hbdh	209.295 ± 57.826	175.906 ± 48.18	Mann-Whitney U	0	6107843
anion_gap	13.026 ± 1.456	12.345 ± 1.368	Mann-Whitney U	0	6496800
ldl	2.686 ± 0.372	2.685 ± 0.361	Mann-Whitney U	0.23394	10140916.5

tt	16.636 ± 0.809	16.432 ± 0.615	Mann-Whitney U	0	7950954.5
nihss	11.529 ± 2.564	7.886 ± 2.871	Mann-Whitney U	0	2984725.5
albumin	40.734 ± 2.37	40.886 ± 2.257	Mann-Whitney U	0.025821	10338834.5
inr	1.068 ± 0.072	1.076 ± 0.149	Mann-Whitney U	0	9016933.5
tg	1.662 ± 0.484	1.536 ± 0.433	Mann-Whitney U	0	7582690.5
bilirubin	16.516 ± 4.009	15.197 ± 3.981	Mann-Whitney U	0	7522775
ima	81.624 ± 8.559	75.458 ± 12.891	Mann-Whitney U	0	4487861
pt	13.822 ± 0.627	13.843 ± 1.151	Mann-Whitney U	0	8374380.5
crp	55.681 ± 48.823	15.314 ± 18.865	Mann-Whitney U	0	3060302
wbc	11.79 ± 3.084	8.316 ± 1.286	Mann-Whitney U	0	2667973
age	65.335 ± 13.909	66.806 ± 12.597	Mann-Whitney U	0.013188	10386092
hdl	1.246 ± 0.146	1.249 ± 0.149	Mann-Whitney U	0.619502	10008026
lactate	2.825 ± 0.376	2.505 ± 0.411	Mann-Whitney U	0	4480425
rbc	4.408 ± 0.274	4.304 ± 0.324	Mann-Whitney U	0	7811417
ast	38.25 ± 18.205	26.05 ± 12.823	Mann-Whitney U	0	3814876
plt	180.251 ± 36.939	190.132 ± 26.424	Mann-Whitney U	0	11826502.5
alt	26.827 ± 10.349	24.193 ± 10.108	Mann-Whitney U	0	7632233.5

aptt	35.045 ± 1.881	35.702 ± 2.313	Mann-Whitney U	0	11737054.5
ldh	296.455 ± 111.282	215.357 ± 75.036	Mann-Whitney U	0	5261997.5
creatinine	83.837 ± 24.574	85.199 ± 52.439	Mann-Whitney U	0	8567930.5
hba1c	6.759 ± 1.048	6.662 ± 0.916	Mann-Whitney U	0.000035	9132523
urea	6.33 ± 1.354	6.419 ± 1.438	Mann-Whitney U	0.001566	10515532
ck	1029.594 ± 872.8	195.007 ± 273.212	Mann-Whitney U	0	3469376

Table 1.Single factor significant analysis results

Feature	0 (N=20789)	1 (N=954)	OR (univariable)	coef	std err	z	P>|z|	[0.025	0.975]	Label_1	Label_0
age	66.806 ± 12.597	65.335 ± 13.909	0.991 (0.986-0.996, p=0.0)	-0.0090	0.003	-3.508	0.000	-0.014	-0.004	-	-
plt	190.132 ± 26.424	180.251 ± 36.939	0.986 (0.983-0.988, p=0.0)	-0.0141	0.001	-11.320	0.000	-0.017	-0.012	-	-
wbc	8.316 ± 1.286	11.79 ± 3.084	2.23 (2.149-2.314, p=0.0)	0.8022	0.019	42.306	0.000	0.765	0.839	-	-
rbc	4.304 ± 0.324	4.408 ± 0.274	2.622 (2.162-3.177, p=0.0)	0.9638	0.098	9.805	0.000	0.771	1.156	-	-
hba1c	6.662 ± 0.916	6.759 ± 1.048	1.112 (1.042-1.186, p=0.001)	0.1059	0.033	3.176	0.001	0.041	0.171	-	-
crp	15.314 ± 18.865	55.681 ± 48.823	1.033 (1.031-1.035,	0.0326	0.001	36.79	0.00	0.031	0.034	-	-

			p=0.0)			2	0				
tg	1.536 ± 0.433	1.662 ± 0.484	1.617 (1.441-1.815, p=0.0)	0.4807	0.059	8.170	0.000	0.365	0.596	-	-
ldl	2.685 ± 0.361	2.686 ± 0.372	1.009 (0.843-1.207, p=0.924)	0.0087	0.091	0.095	0.924	-0.171	0.188	-	-
hdl	1.249 ± 0.149	1.246 ± 0.146	0.87 (0.562-1.349, p=0.534)	-0.1389	0.223	-0.622	0.534	-0.577	0.299	-	-
ast	26.05 ± 12.823	38.25 ± 18.205	1.028 (1.024-1.031, p=0.0)	0.0277	0.002	17.007	0.000	0.024	0.031	-	-
alt	24.193 ± 10.108	26.827 ± 10.349	1.017 (1.012-1.021, p=0.0)	0.0169	0.002	7.507	0.000	0.012	0.021	-	-
bilirubin	15.197 ± 3.981	16.516 ± 4.009	1.068 (1.054-1.082, p=0.0)	0.0662	0.007	9.826	0.000	0.053	0.079	-	-
albumin	40.886 ± 2.257	40.734 ± 2.37	0.971 (0.945-0.999, p=0.042)	-0.0291	0.014	-2.036	0.042	-0.057	-0.001	-	-
urea	6.419 ± 1.438	6.33 ± 1.354	0.955 (0.91-1.002, p=0.063)	-0.0459	0.025	-1.862	0.063	-0.094	0.002	-	-
creatinine	85.199 ± 52.439	83.837 ± 24.574	0.999 (0.998-1.001, p=0.425)	-0.0006	0.001	-0.798	0.425	-0.002	0.001	-	-
bua	344.132 ± 58.336	342.521 ± 74.651	1.0 (0.998-1.001, p=0.411)	-0.0005	0.001	-0.822	0.411	-0.002	0.001	-	-
pt	13.843 ± 1.151	13.822 ± 0.627	0.982 (0.925-1.043, p=0.564)	-0.0177	0.031	-0.577	0.564	-0.078	0.042	-	-

aptt	35.702 ± 2.313	35.045 ± 1.881	0.863 (0.835-0.891, p=0.0)	-0.1473	0.017	-8.917	0.000	-0.180	-0.115	-	-
tt	16.432 ± 0.615	16.636 ± 0.809	1.411 (1.287-1.547, p=0.0)	0.3442	0.047	7.328	0.000	0.252	0.436	-	-
inr	1.076 ± 0.149	1.068 ± 0.072	0.643 (0.385-1.074, p=0.091)	-0.421	0.262	-1.689	0.091	-0.955	0.071	-	-
d_dimer	1.198 ± 0.98	4.362 ± 4.398	1.717 (1.662-1.774, p=0.0)	0.5405	0.017	32.724	0.000	0.508	0.573	-	-
fibrinogen	3.602 ± 0.464	3.518 ± 0.663	0.675 (0.585-0.778, p=0.0)	-0.3931	0.073	-5.408	0.000	-0.536	-0.251	-	-
ck	195.007 ± 273.212	1029.594 ± 872.8	1.002 (1.002-1.002, p=0.0)	0.0024	6.15e-05	38.326	0.000	0.002	0.002	-	-
ldh	215.357 ± 75.036	296.455 ± 111.282	1.005 (1.005-1.006, p=0.0)	0.0053	0.000	21.424	0.000	0.005	0.006	-	-
hbdh	175.906 ± 48.18	209.295 ± 57.826	1.006 (1.005-1.007, p=0.0)	0.0062	0.000	15.637	0.000	0.005	0.007	-	-
ima	75.458 ± 12.891	81.624 ± 8.559	1.015 (1.012-1.017, p=0.0)	0.0147	0.001	10.707	0.000	0.012	0.017	-	-
lactate	2.505 ± 0.411	2.825 ± 0.376	3.12 (2.784-3.494, p=0.0)	1.1377	0.058	19.587	0.000	1.024	1.251	-	-
anion_gap	12.345 ± 1.368	13.026 ± 1.456	1.344 (1.29-1.399, p=0.0)	0.2953	0.021	14.368	0.000	0.255	0.336	-	-
tco2	22.781	22.739	0.972 (0.921-	-0.0	0.0	-1.0	0.	-0.0	0.0	-	-

	± 1.225	± 1.025	1.025, p=0.293)	287	27	51	293	82	25		
nihss	7.886 ± 2.871	11.529 ± 2.564	1.342 (1.318-1.368, p=0.0)	0.2942	0.010	30.957	0.0000	0.276	0.313	-	-
uremia_0	20618 (99.177%)	934 (97.904%)	-	-	-	-	-	-	-	4.334% (934 / 21552)	95.666% (20618 / 21552)
uremia_1	171 (0.823%)	20 (2.096%)	2.582 (1.618-4.121, p=0.0)	0.9485	0.239	3.974	0.000	0.481	1.416	10.471% (20 / 191)	89.529% (171 / 191)
dvt_0	19534 (93.963%)	847 (88.784%)	-	-	-	-	-	-	-	4.156% (847 / 20381)	95.844% (19534 / 20381)
dvt_1	1255 (6.037%)	107 (11.216%)	1.966 (1.595-2.423, p=0.0)	0.6761	0.107	6.340	0.000	0.467	0.885	7.856% (107 / 1362)	92.144% (1255 / 1362)
fatty_liver_0	16655 (80.114%)	812 (85.115%)	-	-	-	-	-	-	-	4.649% (812 / 17467)	95.351% (16655 / 17467)
fatty_liver_1	4134 (19.886%)	142 (14.885%)	0.705 (0.587-0.845, p=0.0)	-0.3502	0.093	-3.782	0.000	-0.532	-0.169	3.321% (142 / 4276)	96.679% (4134 / 4276)
diabetes_0	13737 (66.078%)	617 (64.675%)	-	-	-	-	-	-	-	4.298% (617 / 14354)	95.702% (13737 / 14354)
diabetes_1	7052 (33.922%)	337 (35.325%)	1.064 (0.929-1.219, p=0.371)	0.0620	0.069	0.895	0.371	-0.074	0.198	4.561% (337 / 7389)	95.439% (7052 / 7389)
hypertension_0	6497 (31.252%)	290 (30.398%)	-	-	-	-	-	-	-	4.273% (290 / 6787)	95.727% (6497 / 6787)
hypertension_1	14292 (68.748%)	664 (69.602%)	1.041 (0.904-1.198, p=0.578)	0.0400	0.072	0.556	0.578	-0.101	0.181	4.44% (664 / 14956)	95.56% (14292 / 14956)
coronary_disease_0	11288 (54.298%)	599 (62.788%)	-	-	-	-	-	-	-	5.039% (599 / 11887)	94.961% (11288 / 11887)

coronary_disease_1	9501 (45.702 %)	355 (37.212 %)	0.704 (0.616-0.805, p=0.0)	-0.3508	0.068	-5.128	0.000	-0.485	-0.217	3.602% (355 / 9856)	96.398% (9501 / 9856)
atrial_fibrillation_0	18811 (90.485 %)	838 (87.841 %)	-	-	-	-	-	-	-	4.265% (838 / 19649)	95.735% (18811 / 19649)
atrial_fibrillation_1	1978 (9.515 %)	116 (12.159 %)	1.316 (1.078-1.608, p=0.007)	0.2749	0.102	2.699	0.007	0.075	0.475	5.54% (116 / 2094)	94.46% (1978 / 2094)
hyperuricemia_0	18547 (89.215 %)	801 (83.962 %)	-	-	-	-	-	-	-	4.14% (801 / 19348)	95.86% (18547 / 19348)
hyperuricemia_1	2242 (10.785 %)	153 (16.038 %)	1.58 (1.322-1.889, p=0.0)	0.4575	0.091	5.027	0.000	0.279	0.636	6.388% (153 / 2395)	93.612% (2242 / 2395)
hyperlipidaemia_0	16439 (79.075 %)	801 (83.962 %)	-	-	-	-	-	-	-	4.646% (801 / 17240)	95.354% (16439 / 17240)
hyperlipidaemia_1	4350 (20.925 %)	153 (16.038 %)	0.722 (0.605-0.861, p=0.0)	-0.3259	0.090	-3.627	0.000	-0.502	-0.150	3.398% (153 / 4503)	96.602% (4350 / 4503)
hypoproteinemia_0	18479 (88.888 %)	774 (81.132 %)	-	-	-	-	-	-	-	4.02% (774 / 19253)	95.98% (18479 / 19253)
hypoproteinemia_1	2310 (11.112 %)	180 (18.868 %)	1.86 (1.573-2.201, p=0.0)	0.6208	0.086	7.248	0.000	0.453	0.789	7.229% (180 / 2490)	92.771% (2310 / 2490)
cerebral_hernia_0	20626 (99.216 %)	934 (97.904 %)	-	-	-	-	-	-	-	4.332% (934 / 21560)	95.668% (20626 / 21560)
cerebral_hernia_1	163 (0.784 %)	20 (2.096 %)	2.71 (1.696-4.332, p=0.0)	0.9968	0.239	4.166	0.000	0.528	1.466	10.929% (20 / 183)	89.071% (163 / 183)
hydrocephalus_0	20565 (98.923 %)	895 (93.816 %)	-	-	-	-	-	-	-	4.171% (895 / 21460)	95.829% (20565 / 21460)
hydrocephalus_1	224 (1.077 %)	59 (6.184 %)	6.052 (4.509-8.125,	1.8004	0.150	11.98	0.00	1.506	2.095	20.848% (59 / 283)	79.152% (224 / 283)

			p=0.0)			2	0				
frontal_lobe_0	19943 (95.931 %)	868 (90.985 %)	-	-	-	-	-	-	-	4.171% (868 / 20811)	95.829% (19943 / 20811)
frontal_lobe_1	846 (4.069 %)	86 (9.015 %)	2.336 (1.852-2.945, p=0.0)	0.8483	0.118	7.166	0.0000	0.616	1.080	9.227% (86 / 932)	90.773% (846 / 932)
parietal_lobe_0	20180 (97.071 %)	884 (92.662 %)	-	-	-	-	-	-	-	4.197% (884 / 21064)	95.803% (20180 / 21064)
parietal_lobe_1	609 (2.929 %)	70 (7.338 %)	2.624 (2.03-3.391, p=0.0)	0.9647	0.131	7.375	0.0000	0.708	1.221	10.309% (70 / 679)	89.691% (609 / 679)
temporal_lobe_0	20209 (97.21 %)	886 (92.872 %)	-	-	-	-	-	-	-	4.2% (886 / 21095)	95.8% (20209 / 21095)
temporal_lobe_1	580 (2.79%)	68 (7.128 %)	2.674 (2.063-3.469, p=0.0)	0.9836	0.133	7.413	0.0000	0.724	1.244	10.494% (68 / 648)	89.506% (580 / 648)
occipital_lobe_0	20422 (98.235 %)	919 (96.331 %)	-	-	-	-	-	-	-	4.306% (919 / 21341)	95.694% (20422 / 21341)
occipital_lobe_1	367 (1.765 %)	35 (3.669 %)	2.119 (1.489-3.016, p=0.0)	0.7511	0.180	4.170	0.0000	0.398	1.104	8.706% (35 / 402)	91.294% (367 / 402)
insular_lobe_0	20519 (98.701 %)	938 (98.323 %)	-	-	-	-	-	-	-	4.372% (938 / 21457)	95.628% (20519 / 21457)
insular_lobe_1	270 (1.299 %)	16 (1.677 %)	1.296 (0.78-2.155, p=0.317)	0.2595	0.259	1.000	0.3117	-0.249	0.768	5.594% (16 / 286)	94.406% (270 / 286)
range_lobe_0	19559 (94.083 %)	830 (87.002 %)	-	-	-	-	-	-	-	4.071% (830 / 20389)	95.929% (19559 / 20389)
range_lobe_1	467 (2.246 %)	43 (4.507 %)	2.17 (1.576-2.989, p=0.0)	0.7746	0.163	4.745	0.0000	0.455	1.095	8.431% (43 / 510)	91.569% (467 / 510)
range_lo	329 (1.583	32 (3.354	2.292 (1.584-	0.8	0.1	4.3	0.	0.4	1.1	8.864% (32 /	91.136% (329 /

be_2	%)	%)	3.317, p=0.0)	294	89	99	000	60	99	361)	361)
range_lobe_3	224 (1.077 %)	31 (3.249 %)	3.261 (2.226-4.778, p=0.0)	1.1821	0.195	6.066	0.0000	0.800	1.564	12.157% (31 / 255)	87.843% (224 / 255)
range_lobe_4	175 (0.842 %)	15 (1.572 %)	2.02 (1.186-3.438, p=0.01)	0.7030	0.271	2.591	0.010	0.171	1.235	7.895% (15 / 190)	92.105% (175 / 190)
range_lobe_5	35 (0.168 %)	3 (0.314 %)	2.02 (0.62-6.58, p=0.243)	0.7030	0.603	1.167	0.243	-0.478	1.884	7.895% (3 / 38)	92.105% (35 / 38)
basal_ganglia_0	19869 (95.575 %)	893 (93.606 %)	-	-	-	-	-	-	-	4.301% (893 / 20762)	95.699% (19869 / 20762)
basal_ganglia_1	920 (4.425 %)	61 (6.394 %)	1.475 (1.129-1.927, p=0.004)	0.3888	0.137	2.847	0.004	0.121	0.656	6.218% (61 / 981)	93.782% (920 / 981)
brainstem_0	20532 (98.764 %)	938 (98.323 %)	-	-	-	-	-	-	-	4.369% (938 / 21470)	95.631% (20532 / 21470)
brainstem_1	257 (1.236 %)	16 (1.677 %)	1.363 (0.819-2.268, p=0.234)	0.3095	0.260	1.191	0.234	-0.200	0.819	5.861% (16 / 273)	94.139% (257 / 273)
encephalon_0	20362 (97.946 %)	934 (97.904 %)	-	-	-	-	-	-	-	4.386% (934 / 21296)	95.614% (20362 / 21296)
encephalon_1	427 (2.054 %)	20 (2.096 %)	1.021 (0.649-1.606, p=0.928)	0.0209	0.231	0.090	0.928	-0.432	0.474	4.474% (20 / 447)	95.526% (427 / 447)
paraventricular_0	19786 (95.175 %)	899 (94.235 %)	-	-	-	-	-	-	-	4.346% (899 / 20685)	95.654% (19786 / 20685)
paraventricular_1	1003 (4.825 %)	55 (5.765 %)	1.207 (0.912-1.597, p=0.187)	0.1880	0.143	1.318	0.187	-0.092	0.468	5.198% (55 / 1058)	94.802% (1003 / 1058)
centrum_semiovale_0	20207 (97.2%)	922 (96.646 %)	-	-	-	-	-	-	-	4.364% (922 / 21129)	95.636% (20207 / 21129)

centrum_semiovale_1	582 (2.8%)	32 (3.354%)	1.205 (0.839-1.73, p=0.313)	0.1865	0.185	1.010	0.313	-0.175	0.548	5.212% (32 / 614)	94.788% (582 / 614)
thalamus_0	20565 (98.923%)	937 (98.218%)	-	-	-	-	-	-	-	4.358% (937 / 21502)	95.642% (20565 / 21502)
thalamus_1	224 (1.077%)	17 (1.782%)	1.666 (1.013-2.74, p=0.044)	0.5102	0.254	2.011	0.044	0.013	1.008	7.054% (17 / 241)	92.946% (224 / 241)
aca_0	20524 (98.725%)	941 (98.637%)	-	-	-	-	-	-	-	4.384% (941 / 21465)	95.616% (20524 / 21465)
aca_1	265 (1.275%)	13 (1.363%)	1.07 (0.611-1.874, p=0.813)	0.0676	0.286	0.236	0.813	-0.493	0.628	4.676% (13 / 278)	95.324% (265 / 278)
mca_0	19998 (96.195%)	912 (95.597%)	-	-	-	-	-	-	-	4.362% (912 / 20910)	95.638% (19998 / 20910)
mca_1	791 (3.805%)	42 (4.403%)	1.164 (0.848-1.598, p=0.348)	0.1521	0.162	0.939	0.348	-0.165	0.469	5.042% (42 / 833)	94.958% (791 / 833)
pca_0	20729 (99.711%)	952 (99.79%)	-	-	-	-	-	-	-	4.391% (952 / 21681)	95.609% (20729 / 21681)
pca_1	60 (0.289%)	2 (0.21%)	0.726 (0.177-2.974, p=0.656)	-0.3205	0.720	-0.445	0.656	-1.731	1.090	3.226% (2 / 62)	96.774% (60 / 62)
va_0	20159 (96.97%)	927 (97.17%)	-	-	-	-	-	-	-	4.396% (927 / 21086)	95.604% (20159 / 21086)
va_1	630 (3.03%)	27 (2.83%)	0.932 (0.631-1.377, p=0.724)	-0.0704	0.199	-0.353	0.724	-0.461	0.320	4.11% (27 / 657)	95.89% (630 / 657)
ba_0	20605 (99.115%)	945 (99.057%)	-	-	-	-	-	-	-	4.385% (945 / 21550)	95.615% (20605 / 21550)
ba_1	184 (0.885%)	9 (0.943%)	1.067 (0.544-2.09, p=0.851)	0.0644	0.343	0.188	0.85	-0.608	0.737	4.663% (9 / 193)	95.337% (184 / 193)

							1				
gender_0	10407 (50.06%)	372 (38.994%)	-	-	-	-	-	-	-	3.451% (372 / 10779)	96.549% (10407 / 10779)
gender_1	10382 (49.94%)	582 (61.006%)	1.568 (1.373-1.791, p=0.0)	0.4500	0.068	6.635	0.000	0.317	0.583	5.308% (582 / 10964)	94.692% (10382 / 10964)
cca_plaque_0	16100 (77.445%)	751 (78.721%)	-	-	-	-	-	-	-	4.457% (751 / 16851)	95.543% (16100 / 16851)
cca_plaque_1	4689 (22.555%)	203 (21.279%)	0.928 (0.792-1.088, p=0.356)	-0.0746	0.081	-0.923	0.356	-0.233	0.084	4.15% (203 / 4892)	95.85% (4689 / 4892)
ica_plaque_0	19392 (93.28%)	878 (92.034%)	-	-	-	-	-	-	-	4.332% (878 / 20270)	95.668% (19392 / 20270)
ica_plaque_1	1397 (6.72%)	76 (7.966%)	1.202 (0.945-1.528, p=0.135)	0.1836	0.123	1.496	0.135	-0.057	0.424	5.16% (76 / 1473)	94.84% (1397 / 1473)
eca_plaque_0	20591 (99.048%)	942 (98.742%)	-	-	-	-	-	-	-	4.375% (942 / 21533)	95.625% (20591 / 21533)
eca_plaque_1	198 (0.952%)	12 (1.258%)	1.325 (0.737-2.382, p=0.347)	0.2812	0.299	0.940	0.347	-0.305	0.868	5.714% (12 / 210)	94.286% (198 / 210)
subcortex_lobe_0	18454 (88.768%)	814 (85.325%)	-	-	-	-	-	-	-	4.225% (814 / 19268)	95.775% (18454 / 19268)
subcortex_lobe_1	2335 (11.232%)	140 (14.675%)	1.359 (1.131-1.634, p=0.001)	0.3070	0.094	3.262	0.001	0.123	0.491	5.657% (140 / 2475)	94.343% (2335 / 2475)

Table 2.Univariable logistic regression results

Feature	0 (N=20789)	1 (N=954)	OR (multivariable)	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]

tg	1.536 ± 0.433	1.662 ± 0.484	2.458 (2.069-2.92, p=0.0)	0.899	0.088	10.23	0	0.727	1.071
rbc	4.304 ± 0.324	4.408 ± 0.274	4.731 (3.274-6.837, p=0.0)	1.554	0.188	8.275	0	1.186	1.922
age	66.806 ± 12.597	65.335 ± 13.909	1.012 (1.004-1.021, p=0.003)	0.012	0.004	2.971	0.003	0.004	0.021
ast	26.05 ± 12.823	38.25 ± 18.205	1.048 (1.04-1.055, p=0.0)	0.046	0.004	12.413	0	0.039	0.054
plt	190.132 ± 26.424	180.251 ± 36.939	0.977 (0.973-0.98, p=0.0)	-0.024	0.002	-13.375	0	-0.027	-0.02
alt	24.193 ± 10.108	26.827 ± 10.349	0.953 (0.942-0.964, p=0.0)	-0.048	0.006	-8.177	0	-0.059	-0.036
ima	75.458 ± 12.891	81.624 ± 8.559	1.006 (1.001-1.012, p=0.014)	0.006	0.003	2.453	0.014	0.001	0.012
ldh	215.357 ± 75.036	296.455 ± 111.282	0.984 (0.982-0.987, p=0.0)	-0.016	0.001	-12.992	0	-0.018	-0.014
tt	16.432 ± 0.615	16.636 ± 0.809	1.13 (1.009-1.265, p=0.034)	0.122	0.058	2.116	0.034	0.009	0.235
crp	15.314 ± 18.865	55.681 ± 48.823	1.032 (1.028-1.036, p=0.0)	0.031	0.002	15.585	0	0.027	0.035
wbc	8.316 ± 1.286	11.79 ± 3.084	2.091 (1.985-2.204, p=0.0)	0.738	0.027	27.583	0	0.685	0.79

ck	195.007 ± 273.212	1029.594 ± 872.8	1.001 (1.001-1.001, p=0.0)	0.001	0	7.86	0	0.001	0.001
subcortex_lobe_0	18454 (88.768 %)	814 (85.325 %)	-	-	-	-	-	-	-
subcortex_lobe_1	2335 (11.232 %)	140 (14.675 %)	1.188 (0.827-1.707 ,p=0.352)	0.172	0.185	0.93	0.352	-0.191	0.535
frontal_lobe_0	19943 (95.931 %)	868 (90.985 %)	-	-	-	-	-	-	-
frontal_lobe_1	846 (4.069%)	86 (9.015%)	4.577 (1.381-15.17 ,p=0.013)	1.521	0.611	2.488	0.013	0.323	2.719
cerebral_hernia_0	20626 (99.216 %)	934 (97.904 %)	-	-	-	-	-	-	-
cerebral_hernia_1	163 (0.784%)	20 (2.096%)	0.846 (0.387-1.85 ,p=0.676)	-0.167	0.399	-0.418	0.676	-0.949	0.615
thalamus_0	20565 (98.923 %)	937 (98.218 %)	-	-	-	-	-	-	-
thalamus_1	224 (1.077%)	17 (1.782%)	0.669 (0.327-1.373 ,p=0.273)	-0.401	0.366	-1.095	0.273	-1.119	0.317
occipital_lobe_0	20422 (98.235 %)	919 (96.331 %)	-	-	-	-	-	-	-

occipital_lobe_1	367 (1.765%)	35 (3.669%)	2.172 (0.741-6.368 ,p=0.157)	0.776	0.549	1.414	0.157	-0.3	1.851
coronary_disease_0	11288 (54.298%)	599 (62.788%)	-	-	-	-	-	-	-
coronary_disease_1	9501 (45.702%)	355 (37.212%)	1.408 (1.151-1.724 ,p=0.001)	0.342	0.103	3.322	0.001	0.14	0.545
hypoproteinemia_0	18479 (88.888%)	774 (81.132%)	-	-	-	-	-	-	-
hypoproteinemia_1	2310 (11.112%)	180 (18.868%)	1.183 (0.9-1.554 ,p=0.228)	0.168	0.139	1.206	0.228	-0.105	0.441
parietal_lobe_0	20180 (97.071%)	884 (92.662%)	-	-	-	-	-	-	-
parietal_lobe_1	609 (2.929%)	70 (7.338%)	6.939 (2.253-21.375 ,p=0.001)	1.937	0.574	3.375	0.001	0.812	3.062
hyperuricemia_0	18547 (89.215%)	801 (83.962%)	-	-	-	-	-	-	-
hyperuricemia_1	2242 (10.785%)	153 (16.038%)	0.938 (0.691-1.275 ,p=0.684)	-0.064	0.156	-0.407	0.684	-0.37	0.243
temporal_lobe_0	20209 (97.21%)	886 (92.872%)	-	-	-	-	-	-	-

temporal_lobe_1	580 (2.79%)	68 (7.128%)	5.242 (1.548-17.752 ,p=0.008)	1.657	0.622	2.662	0.008	0.437	2.876
range_lobe_0	19559 (94.083%)	830 (87.002%)	-	-	-	-	-	-	-
range_lobe_1	467 (2.246%)	43 (4.507%)	0.359 (0.111-1.159 ,p=0.087)	-1.025	0.598	-1.713	0.087	-2.197	0.147
range_lobe_2	329 (1.583%)	32 (3.354%)	0.084 (0.01-0.703 ,p=0.022)	-2.479	1.085	-2.285	0.022	-4.605	-0.352
range_lobe_3	224 (1.077%)	31 (3.249%)	0.011 (0.0-0.231 ,p=0.004)	-4.542	1.569	-2.895	0.004	-7.617	-1.467
range_lobe_4	175 (0.842%)	15 (1.572%)	0.001 (0.0-0.057 ,p=0.001)	-6.666	1.943	-3.43	0.001	-10.475	-2.857
range_lobe_5	35 (0.168%)	3 (0.314%)	0.001 (0.0-0.115 ,p=0.004)	-6.586	2.259	-2.915	0.004	-11.014	-2.159
hydrocephalus_0	20565 (98.923%)	895 (93.816%)	-	-	-	-	-	-	-
hydrocephalus_1	224 (1.077%)	59 (6.184%)	3.251 (1.939-5.451 ,p=0.0)	1.179	0.264	4.471	0	0.662	1.696
gender_0	10407 (50.06%)	372 (38.994%)	-	-	-	-	-	-	-

gender_1	10382 (49.94%)	582 (61.006%)	0.572 (0.454-0.72 ,p=0.0)	-0.558	0.117	-4.753	0	-0.789	-0.328
uremia_0	20618 (99.177%)	934 (97.904%)	-	-	-	-	-	-	-
uremia_1	171 (0.823%)	20 (2.096%)	1.979 (1.098-3.564 ,p=0.023)	0.682	0.3	2.273	0.023	0.094	1.271
atrial_fibrillation_0	18811 (90.485%)	838 (87.841%)	-	-	-	-	-	-	-
atrial_fibrillation_1	1978 (9.515%)	116 (12.159%)	1.446 (1.087-1.923 ,p=0.011)	0.369	0.145	2.534	0.011	0.084	0.654
basal_ganglia_0	19869 (95.575%)	893 (93.606%)	-	-	-	-	-	-	-
basal_ganglia_1	920 (4.425%)	61 (6.394%)	1.024 (0.642-1.633 ,p=0.921)	0.024	0.238	0.099	0.921	-0.443	0.49
dvt_0	19534 (93.963%)	847 (88.784%)	-	-	-	-	-	-	-
dvt_1	1255 (6.037%)	107 (11.216%)	1.254 (0.922-1.706 ,p=0.149)	0.227	0.157	1.443	0.149	-0.081	0.534
hyperlipidaemia_0	16439 (79.075%)	801 (83.962%)	-	-	-	-	-	-	-

hyperlipidaemia_1	4350 (20.925 %)	153 (16.038 %)	0.825 (0.646-1.052 ,p=0.121)	-0.193	0.124	-1.552	0.121	-0.437	0.051
fatty_liver_0	16655 (80.114 %)	812 (85.115 %)	-	-	-	-	-	-	-
fatty_liver_1	4134 (19.886 %)	142 (14.885 %)	0.759 (0.59-0.978 ,p=0.033)	-0.275	0.129	-2.135	0.033	-0.528	-0.023

Table 3.Multivariable logistic regression results

Model	AUC	Accuracy	Sensitivity/Recall	Specificity	F1-score	PPV/precision
LR	0.967 | 0.973	0.928 | 0.927	0.920 | 0.929	0.928 | 0.927	0.530 | 0.524	0.373 | 0.365
NB	0.903 | 0.909	0.938 | 0.936	0.634 | 0.662	0.952 | 0.949	0.474 | 0.472	0.378 | 0.367
DT	0.997 | 0.906	0.993 | 0.970	1.000 | 0.836	0.993 | 0.976	0.930 | 0.706	0.870 | 0.610
GB	0.998 | 0.992	0.987 | 0.980	0.976 | 0.900	0.988 | 0.983	0.871 | 0.794	0.786 | 0.711
RF	1.000 | 0.996	0.997 | 0.989	1.000 | 0.883	0.997 | 0.994	0.967 | 0.873	0.936 | 0.864
MLP	0.996 | 0.984	0.977 | 0.972	0.975 | 0.932	0.977 | 0.974	0.790 | 0.744	0.664 | 0.619
XGB	1.000 | 0.996	0.996 | 0.988	1.000 | 0.897	0.996 | 0.992	0.961 | 0.867	0.926 | 0.840
LGBM	1.000 | 0.996	0.997 | 0.989	1.000 | 0.886	0.997 | 0.993	0.970 | 0.869	0.941 | 0.853
KNN	0.997 | 0.955	0.965 | 0.955	0.999 | 0.890	0.964 | 0.958	0.717 | 0.631	0.560 | 0.489

Table 4.The AUC,Accuracy, Sensitivity, Specificity, F1-score, positive predictive value (PPV) and negative predictive value (NPV) of ML models of train|test groups