medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Predictive Models for Secondary Epilepsy in Patients with Acute Ischemic Stroke Within One Year

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

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

2 Bioengineering College of Chongqing University, Chongqing, China

3 Chongqing Key Laboratory of Emergency Medicine

4 Jinfeng Laboratory, Chongqing, China

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

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

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

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

[†]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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Data availability statement

The codes,models,analysis results was uploaded at https://github.com/conanan/lasso-ml . The full dataset 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 Central University basic research young teachers and students research ability promotion sub-project(2023CDJYGRH-ZD06);by Emergency Medicine Chongqing Key Laboratory Talent Innovation and development joint fund project（2024RCCX10）.

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Objective: Post-stroke epilepsy (PSE) is a major complication that worsens both prognosis and quality of life in patients with ischemic stroke. This study aims to develop an interpretable machine learning model to predict PSE using medical records from four hospitals in Chongqing.

Methods: We collected and analyzed medical records, imaging reports, and laboratory test results from 21,459 patients diagnosed with ischemic stroke. Traditional univariable and multivariable statistical analyses were performed to identify key predictive factors. The dataset was divided into a 70% training set and a 30% testing set. To address class imbalance, the Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors was used. Nine widely applied machine learning algorithms were evaluated and compared using relevant prediction metrics. SHAP (SHapley Additive exPlanations) was used to interpret the model, assessing the contributions of different features.

Results: Regression analyses showed that complications such as hydrocephalus, cerebral hernia, and deep vein thrombosis, as well as brain regions (frontal, parietal, and temporal lobes), significantly contributed to PSE. Factors like age, gender, NIH Stroke Scale (NIHSS) scores, and laboratory results such as WBC count and D-dimer levels were associated with a higher risk of PSE. Among the machine learning models, tree-based methods such as Random Forest, XGBoost, and LightGBM demonstrated strong predictive performance, achieving an AUC of 0.99.

Conclusion: Our model successfully predicts PSE risk, with tree-based models showing superior performance. The NIHSS score, WBC count, and D-dimer were identified as the most important predictors.

Introduction

Stroke is the second leading cause of death globally, with an annual mortality of approximately 5.5 million, and it is also the leading cause of disability, accounting for 50% of cases worldwide [1]. Ischemic stroke comprises about 80% of all stroke cases [2][3]. Post-stroke epilepsy (PSE) is a common complication, with studies reporting that 3-30% of stroke patients develop epilepsy, which adversely affects their prognosis and quality of life [4]. PSE can worsen cognitive, psychiatric, and physical impairments already caused by cerebrovascular disease and related conditions [5]. The highest incidence of PSE occurs within the first year after an acute stroke, accounting for nearly half of the cases [2]. Thus, early prediction and intervention for PSE, especially in ischemic strokes, are critical.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Currently, most studies rely on clinical data to build statistical models using survival analysis, Cox regression [2][6], and multiple linear regression [7] to create basic models for PSE prediction. Last year, Lin et al. developed a radiomics-based model that outperformed conventional clinical models in predicting PSE related to intracerebral hemorrhage (ICH). They suggested that a combined radiomics-clinical model could improve the assessment of individual PSE risk after the first occurrence of ICH, facilitating early diagnosis and treatment [8]. However, subsequent research raised concerns about the use of radiomics, indicating a need for further investigation [9]. Overall, research on PSE prediction remains limited, with most studies focusing on specific risk factors [10][11][8][12] and building simple models, without proposing more comprehensive and scientifically robust prediction models.

Machine learning has gained attention as a powerful tool for building medical models due to its ability to process large datasets and complex information. It has been increasingly applied in neuroscience and clinical prediction [13][14][15]. Previous studies have used machine learning to explore post-stroke cognitive impairments [16], predict stroke and myocardial infarction risks in large artery vasculitis patients [14], develop post-stroke depression models based on liver function tests [17], and predict hematoma expansion in traumatic brain injury (TBI) [18]. Machine learning models can automatically manage both linear and complex nonlinear relationships between variables and offer insights into how different factors contribute to the prediction target—something that is difficult for traditional statistical models. However, machine learning requires substantial amounts of data and is prone to overfitting with small sample sizes. The quality and volume of input data are critical for the algorithm to detect underlying patterns and make accurate predictions.

This study aims to identify key risk factors from various features extracted from the clinical records and test data of ischemic stroke patients. Using these features, we will develop a machine learning-based prediction model for PSE. By leveraging early admission data, we seek to automatically predict the likelihood of PSE occurrence and provide guidance for clinical decision-making and patient care.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Result

Filling of missing data

Missing values were filled using a Random Forest (RF) model, handling one feature at a time. The imputed features were: 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, and NIHSS.

Characteristics of study participants

A total of 21,459 patients were included in the study. The training set consisted of 15,021 patients, with a PSE incidence of 4.3%. The test set contained 6,438 patients, also with a 4.3% incidence of PSE. The external validation cohort included 536 patients from three hospitals. The statistical details of the clinical characteristics are presented in Table 1.

Statistical analysis indicated that patients with a higher likelihood of developing PSE had complications such as uremia, a history of DVT, atrial fibrillation, hyperuricemia, cerebral hernia, and hydrocephalus. The affected brain regions included the frontal, parietal, occipital, and temporal lobes, as well as the cortex, subcortex, basal ganglia, and hypothalamus. General characteristics included age, gender, and NIHSS score. Laboratory indicators associated with a higher risk of PSE included WBC count, HbA1C, CRP, triglycerides, AST, ALT, bilirubin, urea, uric acid, APTT, PT, D-dimer, CK, CK-MB, LDH, HBDH, IMA, lactate, and anion gap. Additionally, significant p-values were found for fatty liver, coronary heart disease, hyperlipidemia, and HDL, with low or negative values of these indicators linked to a higher risk of secondary complications. The results of the statistical analyses, as well as the univariate and multivariate regression analyses, are detailed in Tables 1, 2, and 3.

Performance of machine learning models

The relevant performance indicators of the machine learning models are presented in Table 4, while the ROC curves, calibration curve, and decision curve analysis (DCA) are shown in Figure 3. Among all models, tree-based models such as Random Forest (RF), XGBoost, and LightGBM had the highest AUC scores, outperforming other models. Notably, Random Forest had the highest positive predictive value (PPV) at 0.864, which was the most significant metric in our models. Complex machine learning algorithms performed better than traditional logistic regression. The Brier score of the calibration curve was 0.006, and the DCA demonstrated good clinical decision-making benefits, indicating strong practical value. In the external validation cohort, we used RF for predictions, achieving a sensitivity of 0.91 and a PPV of 0.95, confirming the model's strong predictive capability.

Analysis of SHAP risk factors

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4 shows the SHAP (Shapley Additive Explanations) values, individual decision attempts, and overall decision curves. Among general characteristics, females had a higher rate of PSE. A higher NIHSS score was associated with a higher incidence of PSE. Additionally, elevated values of WBC count, D-dimer, CRP, AST, CK-MB, HbA1c, bilirubin, TCO2, and LDH at admission were linked to a greater likelihood of developing PSE. Conversely, lower levels of HBDH, PLT, and APTT were also associated with a higher probability of PSE. The specific brain regions affected did not have a significant individual effect on the overall outcome. Among complications, hypertension was more strongly associated with PSE development, while other conditions, such as coronary heart disease, diabetes, hyperlipidemia, and fatty liver, were less likely to be related to the outcome. We used the force plot of the first patient to illustrate how different features influenced the prediction. In this case, a prolonged APTT time contributed the most to PSE, followed by elevated AST levels, while a low NIHSS score contributed negatively to the final result. The decision plot aggregated model decisions to show how complex models arrived at their predictions.

Discussion

Our study used comprehensive clinical, imaging, and laboratory data from stroke patients to develop a predictive model using machine learning algorithms. This model achieved an AUC score above 0.95, demonstrating more accurate predictions compared to traditional statistical methods. Our research revealed that tree-based ensemble models provided superior predictive performance, especially when handling large datasets with high-dimensional features.

During the modeling process, due to the extreme imbalance between negative and positive samples, we applied the SMOTEENN technique to resample the dataset, improving the performance of the machine learning models. Through SHAP analysis, we conducted interpretability assessments of the model and identified the importance of different features.

In our study, age and NIHSS scores were treated as continuous variables. We found that female patients, older individuals, and those with higher NIHSS scores were more likely to develop PSE, consistent with recent studies. Higher NIHSS scores, indicating more severe strokes, significantly increased the risk of complications, second only to white blood cell (WBC) count and D-dimer in our model [5][19][10][20]. However, there are differing views on the effect of age. Some studies [5][21] suggest that age below 65 is a high-risk factor, which aligns with our findings, while other studies [22] have found that advanced age is the key factor. Yamada et al. [21] also agreed with our study, indicating that female patients have a higher risk of complications. On the other hand, Waafi et al. [10] reported that male patients are 3.325 times more likely to develop complications, which contradicts our findings.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Previous research has shown that patients with diabetes, dyslipidemia, hypertension, depression, or dementia are at higher risk of developing vascular epilepsy [12]. In our study, statistical analysis and multiple machine learning (ML) models examined the relationship between comorbidities and complications. We found 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 divided into five categories: large artery atherosclerosis, cardioembolism, small vessel occlusion, other determined etiology, and undetermined etiology. Patients with multiple comorbidities often fall into the large artery atherosclerosis and cardioembolism categories, which are more clearly defined and easier to treat, resulting in a lower likelihood of epilepsy. In contrast, strokes of undetermined etiology tend to have worse prognoses and are more likely to lead to epilepsy. Among patients with diabetes, higher HbA1c levels indicate poor blood sugar control and a higher risk of complications. Patients with better control of their blood sugar have a lower overall risk of developing complications.

Alain et al. found that cortical infarction is more likely to lead to epilepsy in patients hospitalized with anterior circulation ischemic stroke [23]. Lin et al. found that factors such as cortical involvement and intracerebral hemorrhage volume increase the likelihood of PSE, which is consistent with our findings [8]. Al-Sahli et al. also suggested that cortical brain injury and large-area lesions raise the risk of PSE [5][21]. In our study, statistics showed that both cortical and subcortical involvement increased the likelihood of PSE, but these regions had less influence compared to other features and were not selected in the LASSO regression.

Previous studies have identified acute infection as a risk factor for ischemic stroke [24]. C-reactive protein (CRP) reflects inflammation levels and is an independent prognostic factor [25]. In our study, both regression and SHAP analysis indicated that WBC count had a significant impact among routine blood test parameters, even surpassing the NIHSS score in SHAP analysis. A high WBC count may indicate severe inflammation or infection, as well as increased blood viscosity, making patients more prone to secondary complications. In general, a high red blood cell count and low platelet count also contributed to an increased risk of complications.

A large-scale study on Chinese individuals found a negative correlation between plasma high-density lipoprotein cholesterol (HDL-C) levels and the risk of ischemic stroke, a weak positive correlation between plasma triglyceride (TG) levels and stroke risk, and a strong correlation between plasma low-density lipoprotein cholesterol (LDL-C) and apolipoprotein B levels [26]. High HDL-C levels are linked to better prognosis [27]. Our study aligns with these findings, showing that high LDL-C, low HDL-C, and elevated TG levels are more likely to result in PSE. This can be understood as high cholesterol and triglyceride levels increase blood viscosity and contribute to vascular sclerosis, promoting clot formation [12][28][29]. Higher D-dimer levels indicate more significant brain tissue damage, increasing the likelihood of PSE. In general, lower activated partial thromboplastin time (APTT) and fibrinogen levels are associated with higher PSE risk, while INR, PT, and TT have a smaller impact. Among liver function indicators, aspartate aminotransferase (AST) had the greatest influence on PSE. High AST, low alanine aminotransferase (ALT), and low albumin levels also had some impact. Lingling Ding et al. found that liver enzyme subgroups defined by ALT and AST were linked to higher risks of adverse outcomes [30], which is consistent with our findings.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Studies have also shown that renal function biomarkers such as urinary microalbumin, cystatin C, and creatinine are associated with higher stroke recurrence rates and poorer prognosis [30]. In our study, low urea levels and high uric acid levels had a negative impact [31][32][33]. Our research supports these conclusions. Elevated uric acid levels at admission were positively associated with PSE, although patients with a prior diagnosis of hyperuricemia were less likely to develop epilepsy. Since uric acid acts as a strong antioxidant and has neuroprotective properties [34], patients with normal liver and kidney function and mild hyperuricemia may have greater resilience in emergencies [35][36]. However, excessively high uric acid levels suggest metabolic disorders and poor liver and kidney function, which are linked to a poor prognosis.

When stroke patients are admitted, cardiac enzyme tests are often conducted to rule out myocardial ischemia. However, studies have shown that elevated CK-MB in stroke patients may not be solely heart-related [37]. Cardiac enzymes are important prognostic indicators [38][39] and have been incorporated into stroke scores [40]. Some studies have reported a higher incidence of abnormal serum cardiac enzyme levels in the acute phase of stroke. While the abnormalities are not related to the stroke type, they are associated with stroke severity, with patients exhibiting consciousness disorders having a significantly higher incidence of abnormal cardiac enzymes than those without such disorders [41]. In our study, CK, CK-MB, and IMA in the cardiac enzyme profile had a significant impact and high predictive value, though further research is required to understand the specific mechanisms involved [34].

Although our study incorporated extensive clinical, imaging, and laboratory data to build more accurate prediction models using machine learning algorithms, surpassing traditional statistical methods, there were still several limitations in the modeling process.

While the current study offers valuable insights, the data sample may not be fully representative, and the model's generalizability requires further evaluation. Although the data was collected from multiple tertiary hospitals and includes over 20,000 cases, earlier data was lost due to hospital system upgrades. The dataset mainly reflects patients diagnosed within the past five years and is predominantly from the Chongqing region, which may limit the model's applicability to other geographic areas.

Additionally, the retrospective nature of the study led to the absence of some important predictive indicators. Many potentially valuable features, such as hemorheology, thromboelastography, and hormone levels, were missing and had to be excluded. Including these features could potentially improve the model's accuracy.

To enhance the predictive power of the model, it would be beneficial to incorporate more data beyond baseline patient characteristics. The current analysis primarily used the results from the first examination upon admission, without fully utilizing information from subsequent exams. In future research, recurrent neural networks could be employed to extract features from the entire sequence of examinations more comprehensively.

To strengthen the study further, data standardization should be improved, and the number of cases and key indicators should continue to grow. Additionally, it would be advantageous to explore more advanced scientific methods, such as deep learning, and utilize all available data to improve prediction accuracy.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Materials and methods

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. Data from three external validation centers—Qianjiang Central Hospital, Bishan District People's Hospital, and Yubei District Traditional Chinese Medicine Hospital—were collected between July 2022 and July 2023 to validate and evaluate the model externally. The external validation cohort emphasized collecting positive cases to test the model's ability to identify these cases accurately.

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 to build a multi-modal stroke patient database. 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 District Traditional Chinese Medicine Hospital.

The selection process is outlined in Figure 1. A total of 42,079 records were retrieved from the stroke database, and 24,733 patients were diagnosed with ischemic or lacunar stroke with new onset. Hemorrhagic strokes (4,565), a history of stroke (2,154), TIA (3,570), unclear cause strokes (561), and records with missing essential data (6,496) were excluded. Patients whose seizures might have been caused by other factors (such as brain tumors, intracranial vascular malformations, or traumatic brain injury) (865), those with a seizure history (152), and patients who died in the hospital (1,444) were also excluded. Additionally, patients lost to follow-up (those without outpatient records or unreachable by phone) or who died within three months of the stroke incident (813) were excluded. Finally, 21,459 cases were included in the study.

Data collection

We extracted all relevant records and data from the hospital databases. Using PostgreSQL, we wrote Structured Query Language (SQL) to manage the data as follows:

(1) General Information: This included gender, age, and NIH Stroke Scale (NIHSS) score at admission.

(2) Comorbidities and Complications: These included uremia, previous deep vein thrombosis (DVT), diabetes mellitus, hypertension, coronary atherosclerosis, atrial fibrillation,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

cerebral hernia, hydrocephalus, hypoproteinemia, hyperuricemia, hyperlipidemia, internal carotid stenosis, and common carotid stenosis.

(3) Brain Involvement (CT or MRI records): We recorded involvement of the cortical lobes and subcortical areas, including the frontal, parietal, temporal, occipital, and insular lobes, as well as the basal ganglia, internal capsule, brain stem, cerebellum, periventricular area, centrum semiovale, and thalamus. The extent of cortical involvement (frontal, parietal, temporal, occipital, and insular lobes) was scored, with each lobe contributing 1 point. Similarly, subcortical involvement (basal ganglia, internal capsule, brain stem, periventricular area, thalamus, and cerebellum) was scored with each area contributing 1 point.

(4) Vascular Involvement (CTA, MRA, or DSA records): We recorded the presence of vascular stenosis or occlusion in the anterior cerebral artery (ACA), middle cerebral artery (MCA), posterior cerebral artery (PCA), vertebral artery (VA), and basilar artery (BA).

(5) Key Laboratory Indicators: These included blood lipids such as triglycerides (TG), high-density lipoprotein cholesterol (HDL), and low-density lipoprotein cholesterol (LDL); liver function indicators such as alanine transaminase (ALT), aspartate aminotransferase (AST), bilirubin, and albumin; renal function markers such as urea, blood uric acid (BUA), and creatinine; blood gas parameters such as lactate, anion gap, and total carbon dioxide (TCO2); coagulation markers such as international normalized ratio (INR), prothrombin time (PT), activated partial thromboplastin time (APTT), thrombin time (TT), D-dimer, and fibrinogen; and myocardial enzymes such as creatine kinase (CK), creatine kinase isoenzyme (CK-MB), lactate dehydrogenase (LDH), ischemic modified albumin (IMA), and α-hydroxybutyrate dehydrogenase (HBDH).

Data processing and model building

Processing of Missing Data: We recorded all laboratory indicators from the first set of tests after stroke admission (every stroke patient undergoes routine blood tests, and liver and kidney function assessments). Indicators with more than 10% missing data were excluded. The remaining indicators with missing values were imputed using the random forest algorithm with default parameters. We processed the features in order of missing values, starting with those that had the least missing data (as this requires the least information for imputation). When imputing a feature, missing values in other features were temporarily replaced with 0. After each regression prediction, the predicted value was inserted into the original feature matrix before proceeding to the next feature. Once all features were processed, the dataset was complete.

Distribution of Characteristics: We used univariate analysis to compare the distribution of characteristics between the PSE-negative and PSE-positive groups. The data were then divided into a training set and a test set in a 7:3 ratio.

Processing of Unbalanced Data: Given the low incidence of PSE and the small proportion of positive cases, we augmented the positive data in the training set using the Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTEENN). The SMOTEENN method from the imblearn Python package was applied with default parameters, and a random seed of 42 was set to ensure reproducibility.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Processing of Categorical Data: For categorical variables, we used the one-hot encoding method for transformation. We then applied the LASSO method to the training set to identify the most important features.

Model Building: First, we used LASSO regression to select the 20 most important features. We then employed 9 commonly used machine learning methods, including Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Multi-Layer Perceptron, XGBoost, LightGBM, and K-Nearest Neighbors. Hyperparameters for each model were optimized through grid search to enhance performance. Model evaluation metrics included accuracy, sensitivity, specificity, F1-score, positive predictive value, and negative predictive value. We also generated ROC curves, calibration curves, and decision curves to further assess model performance. An independent external validation dataset was used to evaluate the generalization ability of the selected model. Lastly, we applied the SHAP algorithm to interpret the best-performing model, analyzing the contribution of each feature to the model's predictions and their clinical relevance. Through this process of model development, optimization, and interpretation, we constructed a machine learning model with strong predictive performance and interpretability, offering valuable support for clinical decision-making.

Statistical approach

PostgreSQL v15 (http://www.postgresql.org/) was used to search and extract 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:

The Shapiro-Wilk test was applied to assess the normality of each feature's distribution. For features that did not follow a normal distribution, the Mann-Whitney U test was used to evaluate their significance in relation to the target variable. For features with a normal distribution, the Levene test was performed to evaluate the homogeneity of variances. Features with homogeneous variances were analyzed using the Student's t-test for significance, while those with heterogeneous variances were analyzed using Welch's t-test.

Confidence intervals for AUC values and Brier scores were calculated using 1,000 bootstrap resampling iterations on the datasets. Binary classification thresholds for the predicted probabilities from all models were established using the maximum Youden index derived from the training cohort.

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

All the code used in this study was uploaded to https://github.com/conanan/lasso-ml.

Conclusion

We developed an interpretable machine learning model to predict the risk of post-stroke epilepsy (PSE) in hospitalized patients with ischemic stroke. Using a large dataset of medical

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

records, our artificial intelligence model demonstrates strong predictive performance for PSE. The key predictors identified by the model include NIHSS score, D-dimer levels, lactate levels, and white blood cell count, along with liver function and cardiac enzyme profile indicators. The model's transparency and interpretability can build trust among clinicians and support decision-making. While the results are promising, further prospective studies are necessary to validate the clinical utility of this tool before it can be applied 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[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.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[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.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[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.

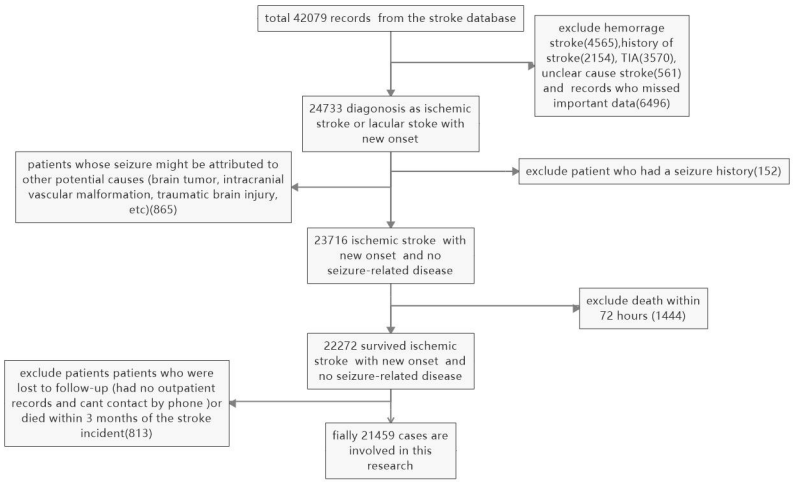
[39] Zeng Y-Y, Zhang W-B, Cheng L, et al.. Cardiac Parameters Affect Prognosis in Patients with Non-Large Atherosclerotic Infarction[J]. Molecular Medicine, 2021, 27(1): 2.

[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.

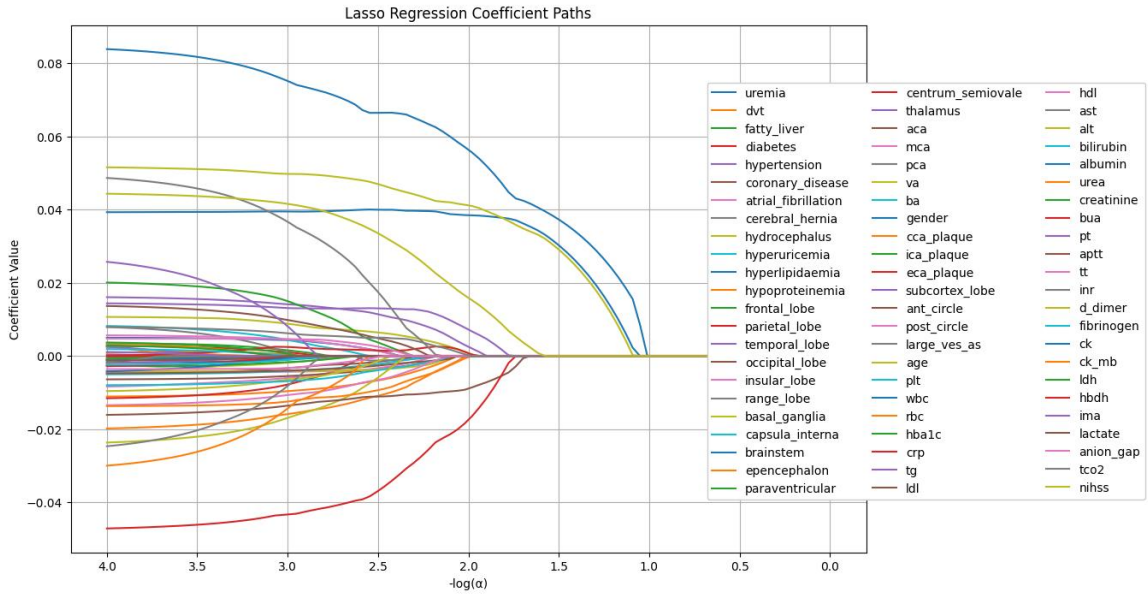
medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 1.Selection and Exclusion Procedure of Patients



A total of 42,079 records were retrieved from the stroke database, and 24,733 patients were diagnosed with ischemic or lacunar stroke with new onset. Hemorrhagic strokes (4,565), a history of stroke (2,154), TIA (3,570), unclear cause strokes (561), and records with missing essential data (6,496) were excluded. Patients whose seizures might have been caused by other factors (such as brain tumors, intracranial vascular malformations, or traumatic brain injury) (865), those with a seizure history (152), and patients who died in the hospital (1,444) were also excluded. Additionally, patients lost to follow-up (those without outpatient records or unreachable by phone) or who died within three months of the stroke incident (813) were excluded. Finally, 21,459 cases were included in the study.

Figure 2.LASSO Regression Coefficient Paths

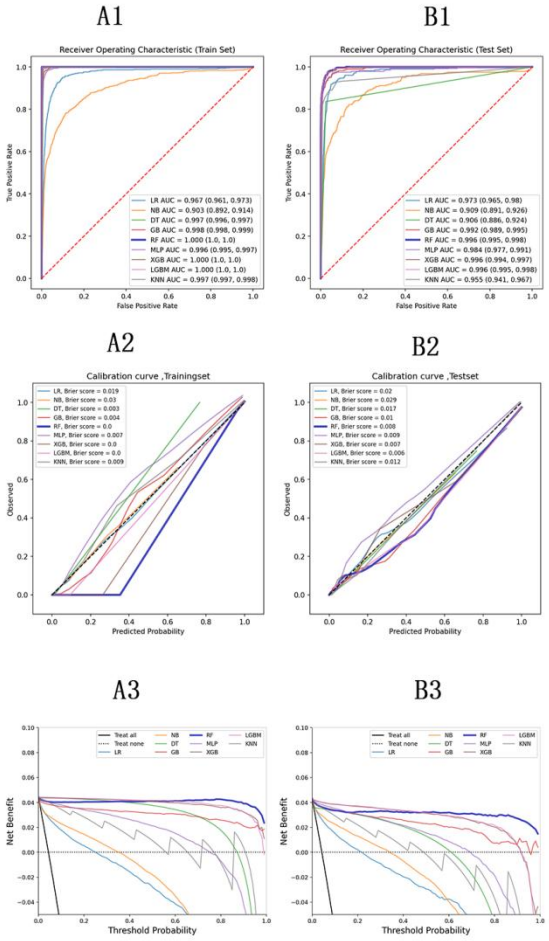


medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The image shows the LASSO regression coefficient paths for various features related to a medical or research study. The x-axis represents the log of the regularization parameter alpha, and the y-axis shows the regression coefficient values.

The lines in the plot represent the coefficient paths for different features as the regularization parameter changes. The features are labeled on the right side of the plot, and the most important features selected by the LASSO model are shown at the bottom of the image.

Figure 3. Model Evaluation Metrics and Curves



The figure displays model performance curves across six sections (A1, A2, A3 on the left; B1, B2, B3 on the right) for training and test sets.

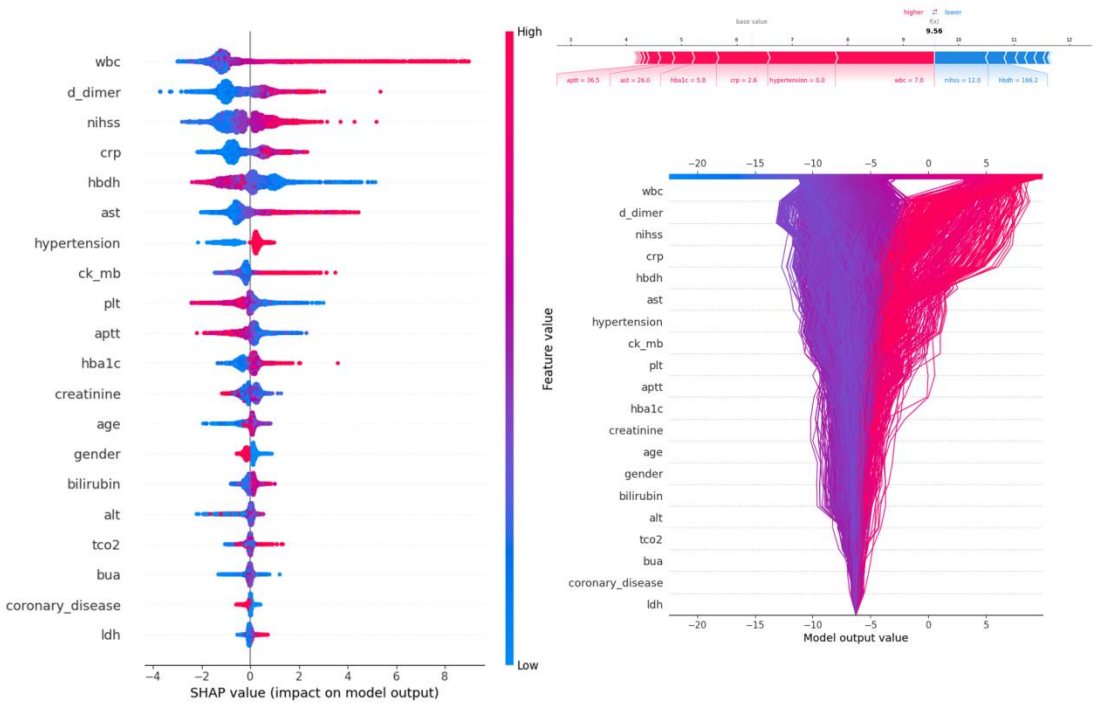
ROC Curve: Illustrates the trade-off between sensitivity and specificity, with the AUC indicating overall model performance.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Calibration Curve: Compares predicted probabilities to actual outcomes, assessing the model's confidence accuracy.

Precision-Recall Curve: Analyzes the balance between precision and recall at various thresholds, particularly useful for imbalanced datasets.

Figure 4.Description of the SHAP Values and Feature Importance



SHAP Value (Left): Displays the impact of each feature on the model's predictions, with features sorted by importance. The color gradient indicates the range of feature values, from low (blue) to high (red).

Force Plot (Upper Right): Illustrates the contribution of individual features of the first sample to the final model output, highlighting how each feature value pushes the prediction away from the baseline value.

Decision Plot (Lower Right): Visualizes the cumulative impact of features on the model output for each sample, showing how the feature values combine to produce the final prediction.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

——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	18479	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	(81.132%)	(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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

——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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

——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%)	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1.Single factor significant analysis results

This table presents the results of Chi-Square and Mann-Whitney U tests used to evaluate the association of various features with positive and negative samples.

Sample Sizes: Positive samples (N=954) and negative samples (N=20,789).
Statistical Methods: The Chi-Square test assesses the relationship between categorical variables, while the Mann-Whitney U test compares differences between independent groups for continuous data.
P-values: Indicate the significance of the associations, with lower values suggesting stronger evidence against the null hypothesis.
Statistical Values: Include counts and percentages of samples for each feature in both groups, along with the calculated statistics for each test.

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	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

crp	15.314 ± 18.865	55.681 ± 48.823	1.033 (1.031-1.035, p=0.0)	0.0326	0.001	36.792	0.000	0.031	0.034	-	-
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 ±	83.837 ±	0.999 (0.998-1.001,	-0.000	0.0	-0.7	0.4	-0.00	0.00	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	52.439	24.574	p=0.425)	6	01	98	25	2	1		
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)	-0.4421	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)	0.053	0.000	21.424	0.000	0.005	0.006	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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 ± 1.225	22.739 ± 1.025	0.972 (0.921-1.025, p=0.293)	-0.0287	0.027	-1.051	0.293	-0.082	0.025	-	-
nihss	7.886 ± 2.871	11.529 ± 2.564	1.342 (1.318-1.368, p=0.0)	0.2942	0.010	30.957	0.000	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)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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_	2242 (10.78	153 (16.03	1.58 (1.322-1.889,	0.457	0.0	5.0	0.0	0.27	0.63	6.388% (153 /	93.612% (2242 /

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1	5%)	8%)	p=0.0)	5	91	27	0 0	9	6	2395)	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, p=0.0)	1.8004	0.150	11.982	0.000	1.506	2.095	20.848% (59 / 283)	79.152% (224 / 283)
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	0.000	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)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

parietal_lobe_1	609 (2.929%)	70 (7.338%)	2.624 (2.03-3.391, p=0.0)	0.9647	0.131	7.375	0.000	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	0.000	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	0.000	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	0.317	-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	0.000	0.455	1.095	8.431% (43 / 510)	91.569% (467 / 510)
range_lobe_2	329 (1.583%)	32 (3.354%)	2.292 (1.584-3.317, p=0.0)	0.8294	0.189	4.399	0.000	0.460	1.199	8.864% (32 / 361)	91.136% (329 / 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	0.000	0.800	1.564	12.157% (31 / 255)	87.843% (224 / 255)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

							0				
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)
epencephalon_0	20362 (97.946%)	934 (97.904%)	-	-	-	-	-	-	-	4.386% (934 / 21296)	95.614% (20362 / 21296)
epencephalon_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_semio	20207 (97.2	922 (96.64	-	-	-	-	-	-	-	4.364% (922 /	95.636% (20207 /

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

vale_0	%)	6%)								21129)	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	0.72	-0.461	0.320	4.11% (27 / 657)	95.89% (630 / 657)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

							4				
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	0.851	-0.608	0.737	4.663% (9 / 193)	95.337% (184 / 193)
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)
subcort	2335	140	1.359							5.657%	94.343%

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ex_lobe_1	(11.232%)	(14.675%)	(1.131-1.634, p=0.001)	0.3070	0.094	3.262	0.001	0.123	0.491	(140 / 2475)	(2335 / 2475)

Table 2.Single Factor Significant Analysis Results

This table presents the results of a single factor significance analysis for various features across two groups of samples: negative samples (0) and positive samples (1).

Sample Size:

Group 0 (Negative): N = 20,789

Group 1 (Positive): N = 954

Feature Analysis: For each feature, the table includes the mean and standard deviation (±) for both groups, odds ratios (OR) from univariable analysis, coefficients (coef), standard errors (std err), z-scores (z), p-values (P>|z|), and 95% confidence intervals ([0.025, 0.975]).

Significance Levels: Features with statistically significant differences are indicated by p-values less than 0.05. An odds ratio greater than 1 suggests an increased risk associated with the feature in the positive group, while an odds ratio less than 1 suggests a decreased risk.

Labels: The last two columns provide the proportions of the positive and negative samples for selected features.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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%)	-	-	-	-	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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 %)	-	-	-	-	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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	-	-	-	-	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	%)	%)							
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	-	-	-	-	-	-	-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	%)	%)						
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 Analysis Results

This table summarizes the results of a multivariable analysis for various features across two groups of samples: negative samples (0) and positive samples (1).

Sample Size:

Group 0 (Negative): N = 20,789

Group 1 (Positive): N = 954

Feature Analysis: For each feature, the table includes the mean and standard deviation (±) for both groups, odds ratios (OR) from multivariable analysis, coefficients (Coef.), standard errors (Std. Err.), z-scores (z), p-values (P>|z|), and 95% confidence intervals ([0.025, 0.975]).

Significance Levels: Features with statistically significant differences are indicated by p-values less than 0.05. An odds ratio greater than 1 indicates an increased risk associated with the feature in the positive group, while an odds ratio less than 1 suggests a decreased risk.

Labels: The last column presents the proportions of the positive and negative samples for selected features.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.27.24309564; this version posted October 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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.Model Performance Evaluation Results
This table presents the performance evaluation metrics for various machine learning models, including AUC, Accuracy, Sensitivity (Recall), Specificity, F1-score, Positive Predictive Value (PPV/Precision), and Negative Predictive Value (NPV)

AUC: Area Under the Curve, indicating the model's ability to distinguish between positive and negative samples; values closer to 1 indicate better performance.

Accuracy: The proportion of correctly classified samples among the total samples.

Sensitivity/Recall: The proportion of correctly identified positive samples out of all actual positive samples.

Specificity: The proportion of correctly identified negative samples out of all actual negative samples.

F1-score: The harmonic mean of precision and recall, considering both the accuracy and completeness of the model.

Positive Predictive Value (PPV/Precision): The proportion of correctly identified positive samples among all samples predicted as positive.

Negative Predictive Value (NPV): The proportion of correctly identified negative samples among all samples predicted as negative.