International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Predicting Stroke-Associated Pneumonia in Acute Ischemic Stroke: A Machine Learning Model Development and Validation Study with CBC-Derived Inflammatory Indices

Mengqi Xie[1,*], Zhiying Liu[1,*], Fangfang Dai[1], Zhen Cao[2], Xiaobei Wang[3,4]

[1]The Second Clinical Medical College of Xinjiang Medical University, Xinjiang Uygur Autonomous Region, People's Republic of China; [2]Department of Clinical Medicine, Xinjiang Medical University, Xinjiang Uygur Autonomous Region, People's Republic of China; [3]Department of Neurology, The Second Affiliated Hospital of Xinjiang Medical University, Xinjiang Uygur Autonomous Region, People's Republic of China; [4]Xinjiang Key Laboratory of Neurological Disorder Research, Xinjiang Uygur Autonomous Region, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xiaobei Wang, Department of Neurology, The Second Affiliated Hospital of Xinjiang Medical University, Xinjiang Uygur Autonomous Region, People's Republic of China, Email wangxiaobei1203@163.com

Purpose: Stroke-associated pneumonia (SAP), a critical complication of ischemic stroke, significantly worsens outcomes. Our aim was to identify SAP risk factors and develop a machine learning (ML) model for early risk stratification.

Methods: This retrospective study analyzed 574 ischemic stroke patients, divided into training (75%) and testing (25%) sets. Nine ML models were trained using 10-fold cross-validation, with performance evaluated by accuracy, AUC-ROC, and F1-score. Key predictors were interpreted via SHAP analysis. An interactive web tool was developed using the optimal model.

Results: SAP incidence was 32.4%. LightGBM demonstrated superior predictive performance (ranking score=54) without overfitting, identifying Monocyte-to-lymphocyte ratio (MLR), systemic immune-inflammation index (SII), NIHSS score, age, aggregate index of systemic inflammation (AISI), and platelet-to-lymphocyte ratio (PLR) as the top predictors.

Conclusion: Our findings demonstrate that machine learning models exhibit strong predictive performance for SAP, with the LightGBM algorithm outperforming other approaches. The web-based prediction tool developed from this model provides clinicians with actionable insights to support real-time clinical decision-making.

Keywords: stroke-associated pneumonia, machine learning, ischemic stroke

Introduction

Stroke-associated infection (SAI) is one of the major complications of stroke and is associated with increased mortality.[1] Studies have shown that approximately 30% of stroke patients develop infections.[2] Among these infections, stroke-associated pneumonia (SAP) is the most severe type of SAI and has the most negative impact on patient prognosis.[3] SAP not only has a high incidence and mortality rate in stroke patients but is also closely related to prolonged hospital stays and poor functional recovery,[4] which significantly affects patient outcomes and rehabilitation. However, prior investigations into SAP preventive strategies and lacks clinical practicability and effectiveness.[5,6] Owing to the absence of validated routine methods in clinical practice to identify patients at the highest risk of SAP, prophylactic antibiotic use often fails.[7] Therefore, accurately identifying patients at risk of SAP in the acute phase of stroke is crucial for implementing preventive strategies and initiating treatment early.

Immune and inflammatory responses play a critical role in the occurrence and progression of ischemic stroke and SAP.[8] Inflammatory-driven neurotoxicity and immune cell cytokine release trigger a counterregulatory anti-inflammatory response, suppressing cytokine production to mitigate infections and halt disease progression. However, persistent inflammatory responses can eventually exhaust the immune system, leading to reduced systemic immune activity, suppression of systemic

Received: 6 March 2025
Accepted: 18 May 2025
Published: 12 June 2025

cellular immune responses, and a rapid decrease in peripheral blood lymphocyte subsets. Therefore, identifying inflammatory biomarkers can be useful for predicting the occurrence of SAP.[9] Among these, inflammatory indices derived from complete blood counts, such as the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), and systemic immune-inflammation index (SII), have been widely used in the diagnosis and prognostic evaluation of various inflammatory diseases.[9–12] These indices are easily obtainable and can reflect the systemic inflammatory status and immune response of the body, making them valuable references in clinical practice.

Machine learning (ML) model is an important branch of artificial intelligence that simulates the human learning process by leveraging vast amounts of data and parallel computing capabilities.[13] It continuously learns from data and self-optimizes to make intelligent decisions. ML models enable more accurate predictive models than traditional statistical methods by analyzing complex clinical data and uncovering latent risk factors associated with disease pathogenesis.[14] In recent years, the application of ML models in the medical field has been increasingly widespread. Studies have shown that ML models hold great potential in the field of stroke, especially in disease prediction, classification accuracy of stroke subtypes, identification of risk factors, and risk assessment.[15–18] However, research on ML-based prediction of complications in acute ischemic stroke is relatively limited.[19–21]Moreover, there are no studies utilizing ML models based on complete blood count-derived inflammatory indices to predict the occurrence of SAP in ischemic stroke patients. This study aims to utilize ML models, combined with inflammatory indices derived from complete blood counts, to construct a predictive model for the risk of SAP in patients with ischemic stroke. By analyzing the clinical characteristics and inflammation-related indicators of patients, this model can provide early warnings for clinicians, enhance the accuracy of SAP risk assessment, assist in formulating personalized treatment plans, and thereby improve patient outcomes. It also promotes the development of personalized prevention strategies and establishes an interactive prediction website to enhance the clinical applicability of the prediction.

Method

To address the research objective of developing a machine learning-based predictive model for assessing SAP risk in ischemic stroke patients, the subsequent sections detail the study's methodological framework. This includes study design, data collection, feature selection strategies, ML models implementation, and evaluation protocols, providing a rigorous foundation for model development and validation.

Study Design and Participants

This retrospective observational study utilized data extracted from the electronic medical record (EMR) system of The Second Affiliated Hospital of Xinjiang Medical University. The study population consisted of ischemic stroke patients admitted to the Department of Neurology between January 2021 and October 2024. Inclusion Criteria: (1) Age > 18 years; (2) Diagnosis of ischemic stroke confirmed according to the Chinese Stroke Association Guidelines for Clinical Management of Ischaemic Cerebrovascular Diseases: Executive Summary and 2023 Update. Exclusion Criteria: (1) Pre-existing pneumonia or other inflammatory conditions prior to admission; (2) History of malignant tumors, hematological disorders, or organ dysfunction; (3) Incomplete clinical data. The study protocol was reviewed and approved by the Ethics Committee of The Second Affiliated Hospital of Xinjiang Medical University (20211012–22C) and conducted in accordance with the Declaration of Helsinki. Verbal informed consent was obtained via telephone follow-up from participants or their legal authorized representatives. A total of 574 patients ultimately met the eligibility criteria for inclusion in the analysis.

Diagnostic Criteria for SAP

All SAP diagnoses were independently validated by three board-certified neurologists following a standardized training protocol. Prior to the initiation of the study, each neurologist completed competency-based training on SAP criteria, including diagnostic simulations requiring a minimum of 90% accuracy for qualification. During formal assessments, two randomly assigned neurologists conducted blinded evaluations, while the third reviewed all discordant cases and 20% of concordant diagnoses for quality control.

The diagnosis of SAP meets the following criteria:[22] At least ONE of the following: (1) Fever (38°C) without other recognized causes. (2) Leukopenia or leukocytosis. (3) Altered mental status (in adults ≥ 70 years) without other recognized causes; At least TWO of the following: (1) New purulent sputum production, change in sputum character within 24 hours, increased respiratory secretions, or increased suctioning requirements. (2) New onset or worsening cough or respiratory rate. (3) Rales, crackles, or bronchial breath sounds. (4) Deterioration in gas exchange; AND ≥ 2 consecutive chest imaging studies demonstrating at least ONE of: n (1) New or progressive and persistent infiltrates. (2) Consolidation. (3) Cavitation.

Collection of Clinical Data

Clinical and laboratory data were extracted from the EMR system: Age, Sex, Body Mass Index (BMI), Hypertension, Diabetes, Coronary disease, National Institutes of Health Stroke Scale (NIHSS) scores at admission, TOAST classification, Smoking, Drinking, and routine blood test results within 24 hours of hospitalization. Based on the 24-hour admission blood tests, we calculated eight inflammatory indices: Monocyte-to-Lymphocyte Ratio (MLR: monocytes/lymphocytes), Neutrophil-to-Lymphocyte Ratio (NLR: neutrophils/lymphocytes), Platelet-to-Lymphocyte Ratio (PLR: platelets/lymphocytes), derived NLR (dNLR: neutrophils/[white blood cells - neutrophils]), Neutrophil-to-(Monocyte + Lymphocyte) Ratio (NMLR: neutrophils/[monocytes + lymphocytes]), Systemic Inflammation Response Index (SIRI: [neutrophils × monocytes]/lymphocytes), Systemic Immune-Inflammation Index (SII: platelets × NLR), and Aggregate Index of Systemic Inflammation (AISI: [neutrophils × monocytes × platelets]/lymphocytes).

Feature Selection in Machine Learning

To address multicollinearity-induced model overfitting, we capitalize on a hybrid feature selection framework integrating unsupervised correlation filtering and supervised ML models. First, unsupervised correlation-based feature elimination was conducted with a stringent threshold (Pearson's $|r| > 0.9$) to remove redundant variables while preserving biological interpretability and critical information. Subsequently, supervised feature selection was implemented through the integrated application of LASSO regression with L1 regularization and Boruta's permutation-based random forest algorithm. This dual approach leverages complementary mechanisms: LASSO achieves dimensionality reduction by enforcing sparsity constraints on linear relationships, whereas Boruta employs shadow feature comparisons to identify non-linear biological interactions through iterative importance assessments. The synergistic combination of these methods enhances model robustness by simultaneously addressing high-dimensional complexity and preserving biologically meaningful patterns. The definitive feature subset was derived from the intersection of variables selected by all three independent selection criteria (correlation filtering, LASSO, and Boruta), ensuring optimal balance between statistical parsimony, biological relevance, and resistance to overfitting through multi-algorithm consensus.

Machine Learning Implementation

Nine ML models [Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), Ridge Regression (RR), k-Nearest Neighbors (KNN), Elastic Net (ENet), Multilayer Perceptron (MLP), Logistic Regression (LR), and Decision Tree (DT)] were constructed using R's tidymodels. This integrative methodology enhances model generalizability. The dataset was split into training (75%) and testing (25%) sets via stratified random sampling. Bayesian hyperparameter optimization with 10-fold cross-validation was applied to the training set, while the testing set was used to evaluated overfitting.

Machine Learning Model Evaluation

For comprehensive model evaluation, we developed a scoring system that integrating eight performance metrics (accuracy, sensitivity, specificity, NPV, PPV, recall, F1-score, and ROC-AUC) across nine candidate models. Each metric's performance ranking was quantified through competitive scoring: models received descending points from 9 (best) to 1 (worst) per metric, with final model selection based on aggregated total scores. This approach objectively synthesizes both discriminative power and clinical decision parameters, while the rank-based weighting inherently normalizes metric heterogeneity and reduces scale-dependent bias.

Feature Importance and Selected Feature Contribution Analysis Using SHAP Values for Optimal Models

The feature importance of the optimal model was evaluated by computing SHapley Additive exPlanations (SHAP) values using the R fastshap package. Global feature rankings were generated based on the mean absolute SHAP values, and the directional contributions of key features were explicitly quantified to interpret their clinical relevance.

Development of a Web-Based Prediction Platform for Real-Time Clinical Assessment

This study developed a web-based interactive Symptom Assessment Prediction (SAP) platform, deploying optimized ML models via the R Shiny framework, with Shiny Server enabling multi-user concurrent access through web browsers.

Statistical Analysis

Data analysis was conducted using R software (version 4.4.1) for ML models implementation and IBM SPSS Statistics (version 27.0) for conventional statistical tests. Continuous variables were expressed as mean ± standard deviation (SD) for normally distributed data (assessed by Shapiro–Wilk test) or median (Q1-Q3) for non-normally distributed data. Group comparisons utilized Student's *t*-test (parametric) or Mann-Whitney-*U* test (non-parametric) for continuous variables and χ^2/Fisher's exact test for categorical variables. ML models were evaluated by accuracy, sensitivity, specificity, PPV, NPV, F1-score, and roc-auc. Statistical significance was defined as two-tailed $P<0.05$.

Results

Building upon the methodological framework established in Section 2, our results demonstrate three key advances: identification of SAP-specific inflammatory signatures through hybrid feature selection, superior predictive capability of LightGBM under multi-criteria evaluation, and clinically actionable interpretation of its outputs via SHAP analysis. The subsequent subsections present these findings in detail.

Baseline Characteristics

Among 574 patients diagnosed with ischemic stroke between January 2021 and October 2024, 186 (32.4%) developed pneumonia. Comparative analysis revealed no statistically significant differences between SAP and non-SAP groups in terms of gender distribution, history of hypertension, diabetes mellitus, smoking status, or alcohol consumption ($P>0.05$). However, statistically significant differences were observed in TOAST classification, age, BMI, NIHSS scores, history of coronary heart disease, and systemic inflammatory indices including NLR, MLR, PLR, dNLR, NMLR. SIRI, SII, and AISI ($P<0.05$) (Table 1).

Feature Selection

During the feature selection process, unsupervised correlation-based filtering initially eliminates highly correlated variables (Pearson's $|r| > 0.9$) via correlation matrix analysis, retaining clinically relevant predictors including Age, BMI, NIHSS, MLR, PLR, dNLR, SII, and AISI (Figure 1A). Subsequently, supervised methods (LASSO regression and the Boruta algorithm) were applied. LASSO regression identified 11 key features: Age, BMI, NIHSS, NLR, MLR, PLR, dNLR, NMLR, SIRI, SII, AISI (Figure 1B and C), while the Boruta algorithm selected 9 biomarkers: Age, NIHSS, NLR, MLR, PLR, NMLR, SIRI, SII, AIS (Figure 1D). Combining results from unsupervised and supervised selection, six consensus features (Age, NIHSS, MLR, PLR, SII, AISI) were ultimately chosen as the core set to develop the final predictive model.

Optimal Hyperparameters of Machine Learning Models

Bayesian optimization combined with 10-fold cross-validation was employed to optimize the hyperparameters of nine ML models, and the optimal hyperparameter sets for these models were ultimately determined, with detailed results presented in Table 2.

Table 1 Comparison of General Information in the Group with SAP and Non-SAP

Characteristics	Total Patients (n =574)	SAP (n =186)	Non-SAP (n =388)	$t/Z/\chi^2$	*P* values
TOAST (n%)[a]				28.982	<0.01
LAA	238(41.46%)	93(50%)	145(37.37%)		
CE	41(7.14%)	24(12.9%)	17(4.38%)		
SAO	273(47.56%)	63(33.87%)	210(54.12%)		
ODC	13(2.26%)	3(1.61%)	10(2.58)		
UND	9(1.57%)	3(1.61%)	6(1.55%)		
Sex (male/female)[a]	391/183	124/62	267/121	0.267	0.605
Age[b]	63 (54,73)	67 (56,79)	61 (53,71)	−4.067	<0.01
BMI[b]	25.01(23.12,27.49)	24.39(22.64,26.94)	25.37 (23.32,27.61)	−2.211	0.027
NIHSS[b]	3(2,6)	4(3,10)	3(2,5)	−5.824	<0.01
Hypertension(n%)[a]	406(70.73%)	130(69.89%)	276(71.13%)	0.094	0.760
Diabetes(n%)[a]	214(37.28%)	74(39.78%)	140(36.08%)	0.737	0.391
Coronary heart (n%)[a]	138(24.04%)	58(31.18%)	80(20.62%)	7.684	0.006
Smoking(n%)[a]	201(35.02%)	58(31.18%)	143(36.86%)	1.778	0.182
Drinking(n%)[a]	144(25.09%)	43(23.12%)	101(26.03%)	0.568	0.451
NLR[b]	2.79(1.92,4.74)	4.18(2.54,6.28)	2.52(1.77,3.67)	−7.362	<0.01
MLR[b]	0.31(0.22,0.44)	0.42(0.26–0.63)	0.27(0.21,0.37)	−7.764	<0.01
PLR[b]	134.28 (103.79,192.06)	162.5(116.06,230.14)	129.08(99.84,173.72)	−5.137	<0.01
dNLR[b]	0.88(0.84,0.91)	0.88(0.84,0.92)	0.87(0.84,0.90)	−2.122	0.034
NMLR[b]	3.12 (2.12,5.20)	4.55(2.88,6.89)	2.83(2.00,4.10)	−7.478	<0.01
SIRI[b]	1.49 (0.87,2.55)	2.16(1.31,4.53)	1.22(0.80, 2.03)	−7.766	<0.01
SII[b]	633.40 (433.34,1108.15)	888.53(553.71,1482.07)	549.12(385.58,875.26)	−6.683	<0.01
AISI[b]	340(192.06,609.61)	477.13(274.17,1077.46)	262.55(168.90,493.41)	−6.826	<0.01

Notes: [a]Categorical variables [n (%)] analyzed by χ^2-test. [b]Non-normally distributed continuous data compared by Mann-Whitney *U*-test.
Abbreviations: NIHSS, National Institutes of Health Stroke Scale; BMI, Body Mass Index, NLR, neutrophil-to-lymphocyte ratio, MLR, monocyte-to-lymphocyte ratio, PLR, platelet-to-lymphocyte ratio, dNLR, derived NLR; NMLR, neutrophil-monocyte-to-lymphocyte ratio; SIRI, systemic inflammation response index; SII, systemic immune-inflammation index; AISI, aggregate index of systemic inflammation; SAP, Stroke-associated pneumonia; TOAST, Trial of Org 10172 in Acute Stroke Treatment classification; LAA, Large Artery Atherosclerosis; CE, Cardioembolism; SAO, Small Artery Occlusion; ODC, Other Determined Cause; UND, Undetermined Cause.

Model Development and Evaluation

We developed nine ML models to predict the occurrence of SAP, including RF, LightGBM, SVM, RR, KNN, ENet, MLP, LR and DT. To evaluate the performance of these nine models, we employed multiple metrics: accuracy, specificity, sensitivity, PPV, NPV, recall, F1-score, and roc-auc. All models demonstrated satisfactory predictive performance on both training and test sets without signs of overfitting, as detailed in Table 3. To identify the optimal model, we implemented a ranking-based scoring system that assessed the performance metrics of all models using the train set (Figure 2). LightGBM emerged as the superior model with the highest composite score of 54 points, establishing it as the most effective predictive model.

Explaining Optimal Models with SHAP Feature Importance

In this study, we utilized the SHAP method to analyze the feature importance of our best-performing model, the LightGBM model. The results indicated that the importance of the six features ranked in the following order: MLR, SII, NIHSS, Age, AISI, and PLR (Figure 3A and B). Further SHAP analysis was conducted to evaluate the effect of these six features within the model, with the x-axis representing the feature values and the y-axis representing the SHAP values. Each point represents the SHAP value and corresponding feature value for each sample, with the sample colors indicating the occurrence of SAP. Results indicate positive associations between MLR, SII, NIHSS, and age with SAP risk, where higher values of these features correlate with increased risk. In contrast, AISI and PLR exhibit different patterns. Unlike traditional inflammatory biomarkers, an elevated AISI is associated with a reduced risk of SAP. We hypothesize that this relationship may be attributed to the platelet-dominated composition of AISI, where platelets play a direct antimicrobial role during the

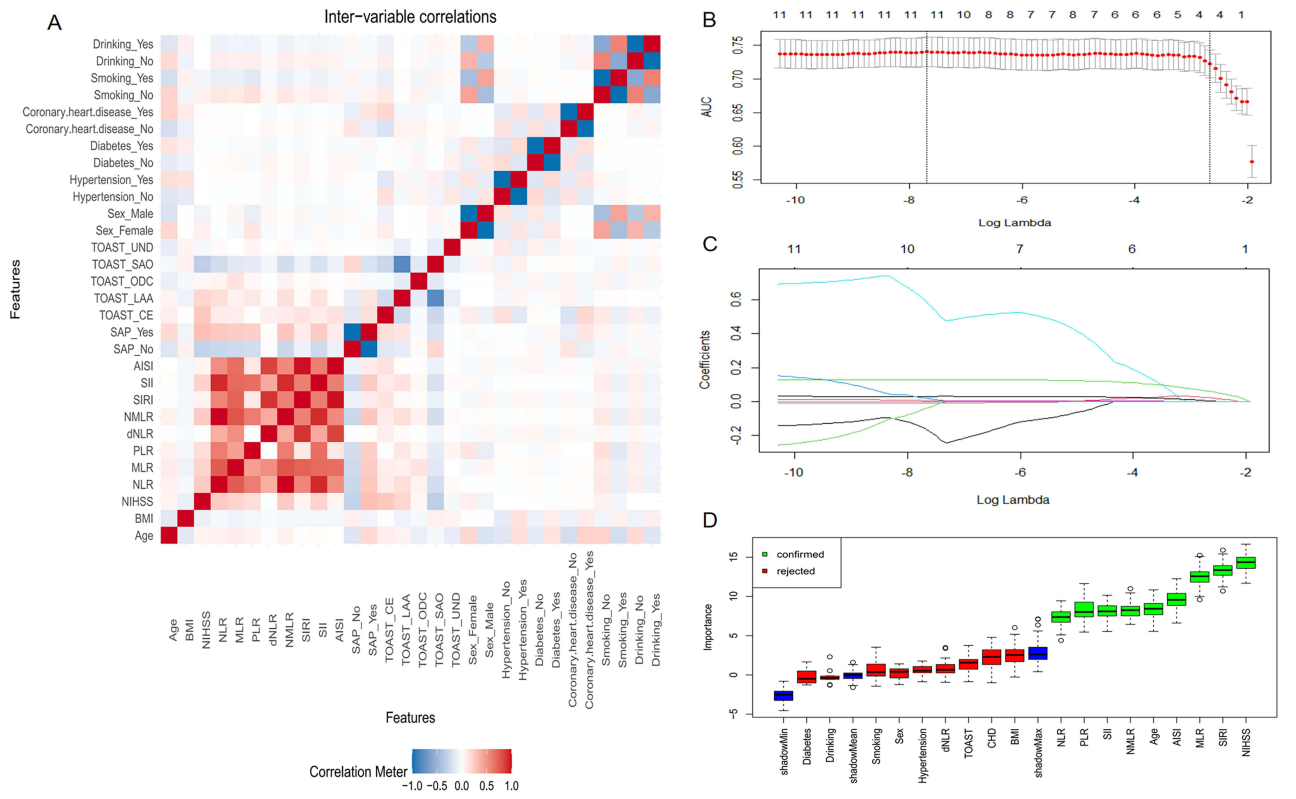

Figure 1 This figure outlines variable selection methodology: (**A**) Heat map visualizing variable correlations through a color gradient. Red hues indicate positive correlations, blue hues represent negative correlations, with color intensity scaled to correlation magnitude (darker shades denote stronger associations). The x-axis and y-axis display clinical variables: Age, BMI, NIHSS (National Institutes of Health Stroke Scale), NLR (neutrophil-to-lymphocyte ratio), MLR (monocyte-to-lymphocyte ratio), PLR (platelet-to-lymphocyte ratio), dNLR (derived NLR), NMLR (neutrophil-monocyte-to-lymphocyte ratio), SIRI (systemic inflammation response index), SII (systemic immune-inflammation index), AISI (aggregate index of systemic inflammation), SAP (Stroke-associated pneumonia), TOAST (Trial of Org 10172 in Acute Stroke Treatment classification), Sex, Hypertension, Diabetes, Coronary heart disease, Smoking, Drinking. (**B**) Optimization of regularization parameter (lambda, λ) through cross-validation Area Under the Curve (AUC) analysis. The x-axis represents the logarithmically transformed regularization parameter [log (λ)], while the y-axis indicates the AUC values. The peak AUC value (marked by red vertical line) identifies the optimal λ that balances model complexity and predictive performance. (**C**) Regularization path tracking coefficient evolution across λ values (y-axis: coefficients); features with coefficients reaching zero are eliminated. (**D**) Feature selection by the Boruta algorithm identifying significant clinical variables (green: confirmed predictors; red: rejected non-contributors). The x-axis displays clinical variables: Age, BMI, NIHSS (National Institutes of Health Stroke Scale), NLR (neutrophil-to-lymphocyte ratio), MLR (monocyte-to-lymphocyte ratio), PLR (platelet-to-lymphocyte ratio), dNLR (derived NLR), NMLR (neutrophil-monocyte-to-lymphocyte ratio), SIRI (systemic inflammation response index), SII (systemic immune-inflammation index), AISI (aggregate index of systemic inflammation), SAP (Stroke-associated pneumonia), TOAST (Trial of Org 10172 in Acute Stroke Treatment classification), Sex, Hypertension, Diabetes, Coronary heart disease, Smoking, Drinking.

hyperacute phase of stroke, before the onset of immune suppression. PLR's relationship with SAP risk is nonlinear. Moderately elevated PLR levels have a weak protective effect, possibly reflecting platelet-mediated pathogen clearance. Both low and high PLR levels contribute to increased SAP risk, indicating a threshold effect in balancing pro-inflammatory and anti-inflammatory responses (Figure 3C). This suggests that PLR might act as a double-edged sword, with moderate levels providing some protection against SAP, whereas both lower and higher levels exacerbate the risk.

Implementation of a LightGBM Mode by Shiny Web Application

Based on the optimal machine learning model-LightGBM, an interactive prediction website was developed with utilizing the Shiny package 0in R (https://prediction-x.shinyapps.io/ascdd/). This platform allows physicians to input patients' relevant clinical feature data, generating personalized prediction results in real time. With this tool, clinicians can more efficiently evaluate patients' conditions, aid in the formulating of treatment plans, and enhance the scientific accuracy and precision of clinical decision-making.

Table 2 Hyperparameter Configuration of Machine Learning Models

Model	Hyperparameters	Hyperparameter Selection
RF	Engine	Randomforest
	mtry	5
	Trees	1936
	min_n	167
LightGBM	Engine	Lightgbm
	Trees	2100
	mtry	2
	min_n	148
	Learn_rate	0.0113
SVM	Engine	Kernlab
	Cost	0.172
	rbf_sigma	0.0508
RR	Engine	glmnet
	Penalty	1.00
KNN	Engine	kknn
	Neighbors	95
ENet	Engine	glmnet
	Mixture	0.109
	Penalty	0.198
MLP	Engine	nnet
	Hidden_units	6
	Penalty	0.973
	Epochs	884
LR	Engine	glm
DT	Engine	rpart
	min_n	132
	Tree_depth	11
	Cost_complexity	0.0817

Abbreviations: RF, Random Forest; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine; RR, Ridge Regression; KNN, k-Nearest Neighbors; ENet, Elastic Net; MLP, Multilayer Perceptron; LR, Logistic Regression and DT, Decision Tree; PPV, positive predictive values; NPV, negative predictive values; roc-auc, area under the ROC curve; min_n, Minimum Number of Observations in a Node for Splitting; rbf_sigma, Radial Basis Function Kernel Sigma; glmnet, Generalized Linear Models with Lasso and Elastic-Net Regularization; kknn, Weighted k-Nearest Neighbors; nnet, Neural Network; rpart, Recursive Partitioning and Regression Trees; glm, Generalized Linear Model.

Table 3 Performance Metrics of the Nine Machine Learning Models

Model		Accuracy	Sensitivity	Specificity	PPV	NPV	Recall	F1-Score	roc-auc
RF	Train set	0.744	0.669	0.780	0.592	0.832	0.669	0.628	0.775
	Test set	0.694	0.596	0.742	0.528	0.791	0.596	0.560	0.728
LightGBM	Train set	0.763	0.612	0.835	0.639	0.818	0.612	0.625	0.769
	Test set	0.750	0.489	0.876	0.657	0.780	0.486	0.561	0.702
SVM	Train set	0.758	0.504	0.880	0.667	0.788	0.504	0.574	0.743
	Test set	0.715	0.340	0.897	0.615	0.737	0.340	0.438	0.700
RR	Train set	0.751	0.468	0.887	0.663	0.777	0.468	0.549	0.733
	Test set	0.722	0.319	0.918	0.652	0.736	0.319	0.429	0.695

(*Continued*)

Table 3 (Continued).

Model		Accuracy	Sensitivity	Specificity	PPV	NPV	Recall	F1-Score	roc-auc
KNN	Train set	0.747	0.468	0.880	0.650	0.776	0.468	0.544	0.720
	Test set	0.729	0.319	0.928	0.682	0.738	0.319	0.435	0.694
ENet	Train set	0.760	0.453	0.907	0.700	0.776	0.453	0.550	0.732
	Test set	0.722	0.298	0.928	0.667	0.732	0.298	0.412	0.683
MLP	Train set	0.770	0.460	0.918	0.727	0.781	0.460	0.564	0.730
	Test set	0.715	0.277	0.928	0.650	0.726	0.277	0.388	0.676
LR	Train set	0.758	0.475	0.893	0.680	0.781	0.475	0.559	0.732
	Test set	0.708	0.298	0.907	0.609	0.727	0.298	0.400	0.674
DT	Train set	0.714	0.597	0.770	0.553	0.800	0.597	0.574	0.683
	Test set	0.708	0.574	0.773	0.551	0.789	0.574	0.562	0.674

Abbreviations: RF, Random Forest; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine; RR, Ridge Regression; KNN, k-Nearest Neighbors; ENet, Elastic Net; MLP, Multilayer Perceptron; LR, Logistic Regression; DT, Decision Tree; PPV, positive predictive values; NPV, negative predictive values; roc-auc, area under the ROC curve.

Discussion

Nine ML models were developed in this study. The LightGBM model was found to have the best predictive performance through comprehensive performance evaluation. Based on this model, we further conducted feature importance analysis using the SHAP method. The results indicated that six features, including MLR, SII, NIHSS, age, AISI, and PLR.

Further analysis revealed that advanced age, high NIHSS score, high SII, and high MLR promote the occurrence of SAP in patients, which is consistent with previous studies.[23–27]Advanced age independently increases the susceptibility

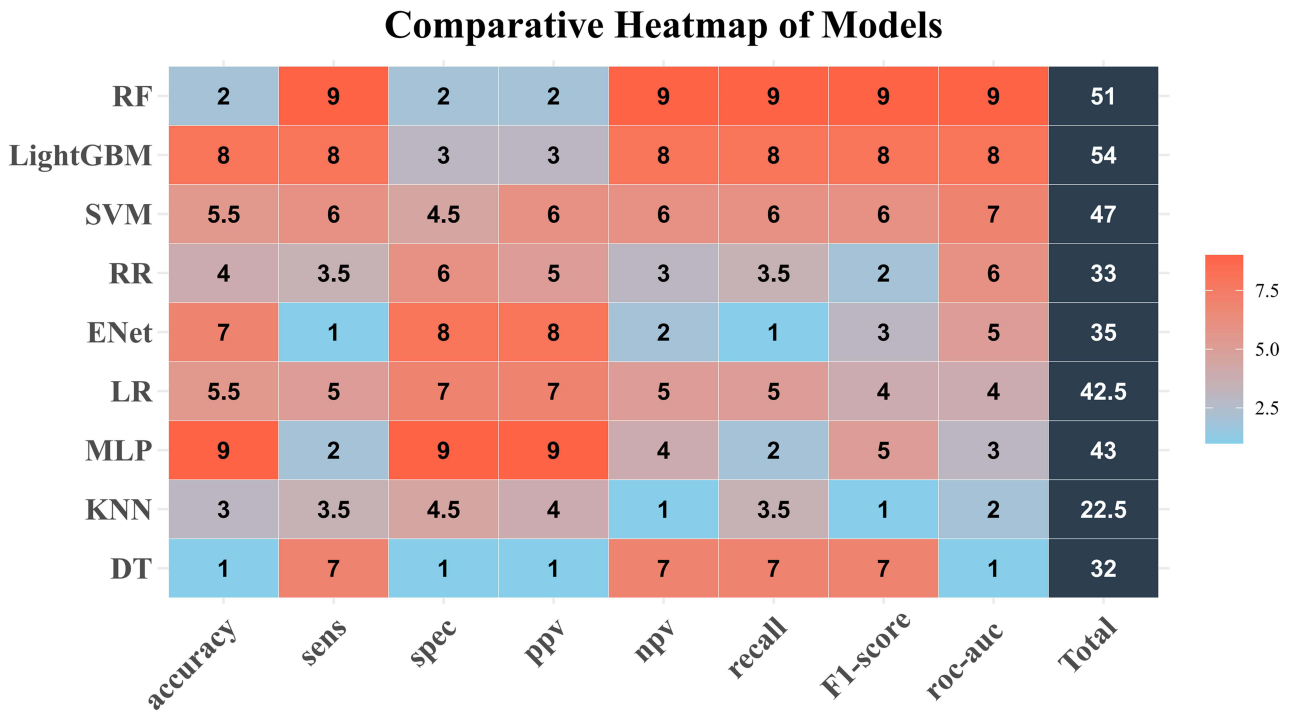

Figure 2 Comparative Heatmap of Models. The rows represent different models: [Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), Ridge Regression (RR), k-Nearest Neighbors (KNN), Elastic Net (ENet), Multilayer Perceptron (MLP), Logistic Regression (LR), and Decision Tree (DT)]. The columns represent different performance metrics: accuracy, sensitivity, specificity, positive/negative predictive values (PPV/NPV), recall, F1-score, and area under the ROC curve (roc-auc). Each cell in the heatmap is color-coded to indicate the performance score, with darker colors representing higher values. The numbers in the cells represent the specific scores for each model-metric combination. The bottom row provides the total scores for each model across all metrics, and the rightmost column provides the average scores for each metric across all models. This heatmap allows for a quick visual assessment of the relative performance of different models across various evaluation criteria.

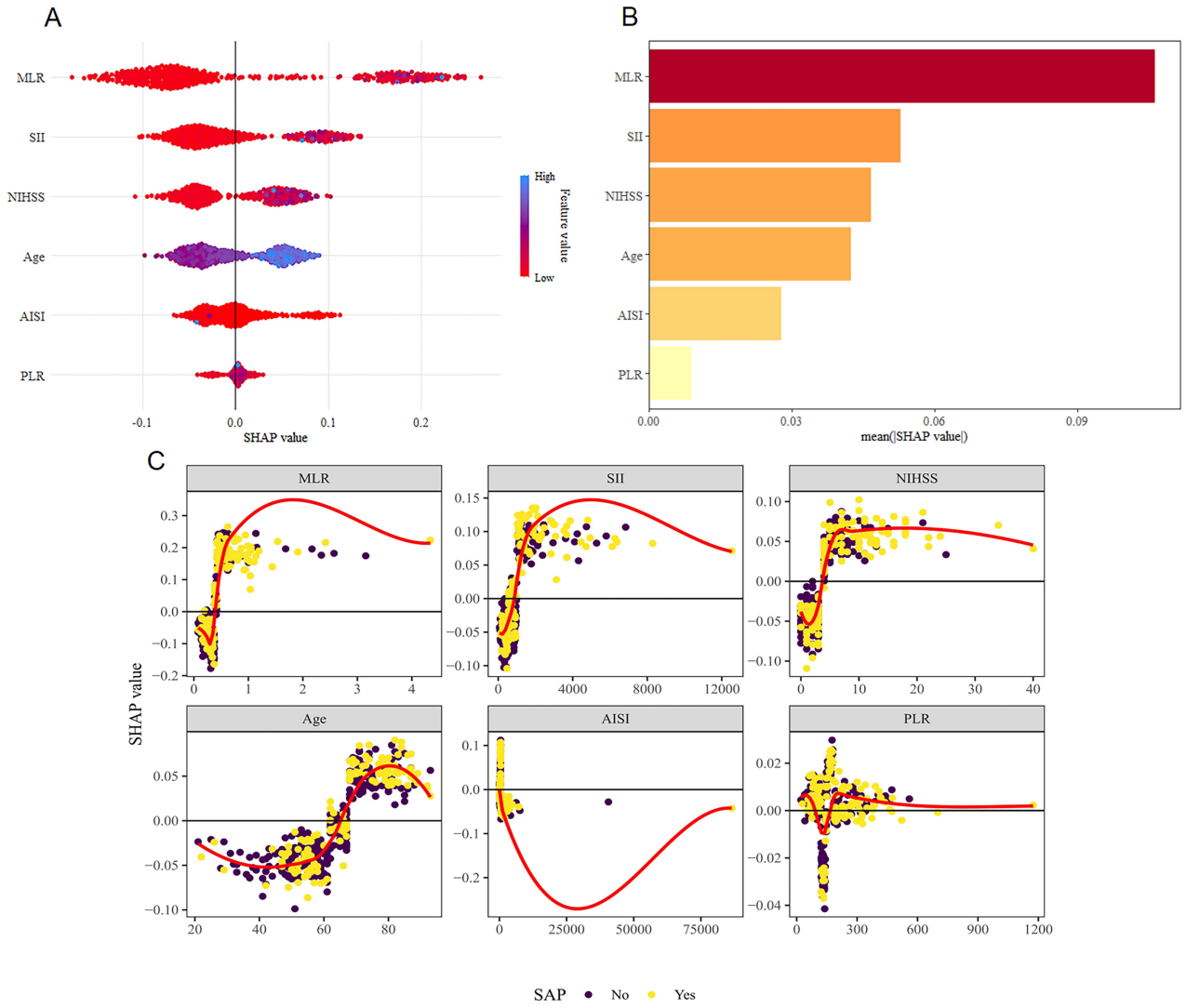

Figure 3 This figure provides a comprehensive analysis of feature importance and their impact on the model output using SHAP values. MLR stands for Monocyte-to-lymphocyte ratio, SII stands for Systemic immune-inflammation index, AISI stands for Aggregate index of systemic inflammation, PLR stands for Platelet-to-lymphocyte ratio. (**A**) It shows the distribution of SHAP values for each feature. (**B**) It displays the mean absolute SHAP values indicating feature importance.(**C**) It illustrates the relationship between feature values and their corresponding SHAP values, differentiated by the presence or absence of SAP.

to post-stroke infection due to immune senescence—age-related declines in adaptive and innate immune responses, often characterized by impaired neutrophil function, reduced T-cell diversity, and weakened mucosal barrier integrity, all of which facilitate bacterial colonization and systemic spread.[28] The NIHSS score reflects the degree of neurological impairment, and a higher NIHSS score indicates greater neurological damage, predisposing patients to aspiration (due to dysphagia or decreased consciousness) and immobility (increasing the risk of atelectasis and ventilator-associated pneumonia).[26,29] Elevated MLR may be related to monocytes. Monocytes are key mediators of innate immunity and can produce pro-inflammatory cytokines (such as IL-6, TNF-α), which can exacerbate tissue damage after stroke. Lymphocytes, particularly T cells and B cells, are crucial for adaptive immunity and pathogen clearance.[30] High SII reflects a high inflammatory state characterized by neutrophilia (acute-phase response), thrombocytosis (platelet activation), and lymphopenia (stroke-induced immunosuppression). Neutrophils release extracellular traps (NETs) that capture pathogens but also damage host tissues.[31] Platelets synergize with neutrophils to amplify NETosis,[32,33] while lymphopenia weakens adaptive immunity.[34] However, AISI showed an unexpected inverse relationship with SAP risk in our analysis. This finding may stem from the transient antimicrobial effects of platelets in the hyperacute phase of stroke.

https://doi.org/10.2147/IJGM.S524450

Platelets can release antimicrobial peptides and enhance phagocytosis,[35] potentially suppressing early bacterial invasion before the onset of immunosuppression. Our study focused only on laboratory test results within 24 hours and did not observe long-term changes in patients' blood cell counts. Therefore, the timing of our measurements and the timing of patient testing may have coincided with the transient "protective inflammation" window in the hyperacute phase. Finally, PLR exhibited a U-shaped relationship with SAP risk: moderate levels provided weak protection, while extreme levels (low or high) increased the risk. Mechanistically, moderately elevated PLR may reflect physiological compensation through the release of antimicrobial factors like thrombopoietin by platelets and promotion of pathogen clearance, whereas low PLR indicates impaired defense capabilities and high PLR signifies excessive inflammation or immune exhaustion.[36,37]

This study has the following limitations: First, the study design is a single-center retrospective analysis, and its external validity needs to be confirmed through multicenter validation. It is also difficult to establish a clear causal relationship. Second, due to the small sample size, there may be an increased risk of selection bias, which affects the generalizability of the conclusions. Third, laboratory indicators were only collected at a single time point without dynamic monitoring. Future studies could consider repeated testing of indicators such as AISI/PLR at different time points after stroke (such as 24 hours,48 hours,72 hours,7 days) to construct a temporal association curve between these indicators and the occurrence and development of SAP. Notably, the methodological framework exhibits transferability to investigations of other types of stroke and related complications in the future. To further enhance clinical translational value, large-scale, multicenter prospective cohort studies are urgently needed to systematically validate the predictive performance and clinical application potential of the model, thereby providing more reliable evidence-based basis for optimizing the whole process management of stroke patients.

Conclusions

The LightGBM model predicts stroke-associated pneumonia utilizing clinical (age, NIHSS) and hematological biomarkers (MLR/SII/AISI/PLR). SHAP analysis was used to further elucidate the magnitude of feature contributions. Deployed as a web tool, it enables real-time risk stratification. This framework refines inflammatory paradigms and enhances precision stroke care by integrating biomarker-pathophysiology insight.

Data Sharing Statement

The datasets used and/or analyzed in this study are available from the corresponding author upon reasonable request.

Ethical

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and its subsequent amendments. The research protocol, which including the use of verbal informed consent, was reviewed and approved by the Ethics Committee of The Second Affiliated Hospital of Xinjiang Medical University (20211012-22C).

Acknowledgments

The authors extend their sincere gratitude to all the patients who participated in this study, as well as to all the staff members who provided assistance in the recruitment of patients and the collection, management, and processing of samples and data. We sincerely thank Ms. Laidandan for her professional English language editing assistance, which significantly improved the clarity and academic rigor of the manuscript.

Funding

This study was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region Youth Project (2021D01C376) and the Undergraduate Innovation and Entrepreneurship Training Program of Xinjiang Medical University (X202310760108).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Wang YJ, Li ZX, Gu HQ, et al. China stroke statistics: an update on the 2019 report from the national center for healthcare quality management in neurological diseases, China national clinical research center for neurological diseases, the Chinese stroke association, national center for chronic and non-communicable disease control and prevention, Chinese center for disease control and prevention and institute for global neuroscience and stroke collaborations. *Stroke Vasc Neurol*. 2022;7(5):415–450. doi:10.1136/svn-2021-001374
2. Westendorp WF, Nederkoorn PJ, Vermeij JD, et al. Post-stroke infection: a systematic review and meta-analysis. *BMC Neurol*. 2011;11(1):110. doi:10.1186/1471-2377-11-110
3. Bustamante A, Giralt D, García-Berrocoso T, et al. The impact of post-stroke complications on in-hospital mortality depends on stroke severity. *Eur Stroke J*. 2017;2(1):54–63. doi:10.1177/2396987316681872
4. Teh WH, Smith CJ, Barlas RS, et al. Impact of stroke-associated pneumonia on mortality, length of hospitalization, and functional outcome. *Acta Neurol Scand*. 2018;138(4):293–300. doi:10.1111/ane.12956
5. Meisel A. Preventive antibiotic therapy in stroke: passed away?. *Lancet*. 2015;385(9977):1486–1487. doi:10.1016/S0140-6736(15)60076-9
6. Badve MS, Zhou Z, Anderson CS, Hackett ML. Effectiveness and safety of antibiotics for preventing pneumonia and improving outcome after acute stroke: systematic review and meta-analysis. *J Stroke Cerebrovasc Dis*. 2018;27(11):3137–3147. doi:10.1016/j.jstrokecerebrovasdis.2018.07.001
7. Faura J, Bustamante A, Miró-Mur F, Montaner J. Stroke-induced immunosuppression: implications for the prevention and prediction of post-stroke infections. *J Neuroinflammation*. 2021;18(1):127. doi:10.1186/s12974-021-02177-0
8. Hoffmann S, Harms H, Ulm L, et al. Stroke-induced immunodepression and dysphagia independently predict stroke-associated pneumonia - the PREDICT study. *J Cereb Blood Flow Metab*. 2017;37(12):3671–3682. doi:10.1177/0271678X16671964
9. Yan D, Dai C, Xu R, Huang Q, Ren W. Predictive ability of systemic inflammation response index for the risk of pneumonia in patients with acute ischemic stroke. *Gerontology*. 2023;69(2):181–188. doi:10.1159/000524759
10. Wang RH, Wen WX, Jiang ZP, et al. The clinical value of neutrophil-to-lymphocyte ratio (NLR), systemic immune-inflammation index (SII), platelet-to-lymphocyte ratio (PLR) and systemic inflammation response index (SIRI) for predicting the occurrence and severity of pneumonia in patients with intracerebral hemorrhage. *Front Immunol*. 2023;14:1115031. doi:10.3389/fimmu.2023.1115031
11. Curbelo J, Luquero Bueno S, Galván-Román JM, et al. Correction: inflammation biomarkers in blood as mortality predictors in community-acquired pneumonia admitted patients: importance of comparison with neutrophil count percentage or neutrophil-lymphocyte ratio. *PLoS One*. 2019;14(2):e0212915. doi:10.1371/journal.pone.0212915
12. Luo S, Yang WS, Shen YQ, et al. The clinical value of neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and D-dimer-to-fibrinogen ratio for predicting pneumonia and poor outcomes in patients with acute intracerebral hemorrhage. *Front Immunol*. 2022;13:1037255. doi:10.3389/fimmu.2022.1037255
13. Sirsat MS, Fermé E, Câmara J. Machine learning for brain stroke: a review. *J Stroke Cerebrovasc Dis*. 2020;29(10):105162. doi:10.1016/j.jstrokecerebrovasdis.2020.105162
14. Rosário M, Fonseca AC. Update on biomarkers associated with large-artery atherosclerosis stroke. *Biomolecules*. 2023;13(8):1251. doi:10.3390/biom13081251
15. Liu Y, Yu Y, Ouyang J, et al. Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model. *Stroke*. 2023;54(9):2316–2327. doi:10.1161/STROKEAHA.123.044072
16. Xie Y, Jiang B, Gong E, et al. JOURNAL CLUB: use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *AJR Am J Roentgenol*. 2019;212(1):44–51. doi:10.2214/AJR.18.20260
17. Brugnara G, Neuberger U, Mahmutoglu MA, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke*. 2020;51(12):3541–3551. doi:10.1161/STROKEAHA.120.030287
18. Huang J, Jin W, Duan X, et al. Twenty-eight-day in-hospital mortality prediction for elderly patients with ischemic stroke in the intensive care unit: interpretable machine learning models. *Front Public Health*. 2022;10:1086339. doi:10.3389/fpubh.2022.1086339
19. Issaiy M, Zarei D, Kolahi S, Liebeskind DS. Machine learning and deep learning algorithms in stroke medicine: a systematic review of hemorrhagic transformation prediction models. *J Neurol*. 2024;272(1):37. doi:10.1007/s00415-024-12810-6
20. Zhang J, Kong Z, Hong S, Zhang Z. Machine learning-based model for prediction of post-stroke cognitive impairment in acute ischemic stroke: a cross-sectional study. *Neurol India*. 2024;72(6):1193–1198. doi:10.4103/ni.ni_987_21
21. Ji W, Wang C, Chen H, et al. Predicting post-stroke cognitive impairment using machine learning: a prospective cohort study. *J Stroke Cerebrovasc Dis*. 2023;32(11):107354. doi:10.1016/j.jstrokecerebrovasdis.2023.107354
22. Smith CJ, Kishore AK, Vail A, et al. Diagnosis of stroke-associated pneumonia: recommendations from the pneumonia in stroke consensus group. *Stroke*. 2015;46(8):2335–2340. doi:10.1161/STROKEAHA.115.009617
23. Ji R, Shen H, Pan Y, et al. Risk score to predict hospital-acquired pneumonia after spontaneous intracerebral hemorrhage. *Stroke*. 2014;45(9):2620–2628. doi:10.1161/STROKEAHA.114.005023
24. Li J, Luo H, Chen Y, et al. Comparison of the predictive value of inflammatory biomarkers for the risk of stroke-associated pneumonia in patients with acute ischemic stroke. *Clin Interv Aging*. 2023;18:1477–1490. doi:10.2147/CIA.S425393
25. Li D, Liu Y, Jia Y, et al. Association between malnutrition and stroke-associated pneumonia in patients with ischemic stroke. *BMC Neurol*. 2023;23(1):290. doi:10.1186/s12883-023-03340-1
26. Wu B, Luo H, Li J, et al. The relationship between the Barthel Index and stroke-associated pneumonia in elderly patients and factors of SAP. *BMC Geriatr*. 2024;24(1):829. doi:10.1186/s12877-024-05400-8
27. Zheng F, Gao W, Xiao Y, et al. Systemic inflammatory response index as a predictor of stroke-associated pneumonia in patients with acute ischemic stroke treated by thrombectomy: a retrospective study. *BMC Neurol*. 2024;24(1):287. doi:10.1186/s12883-024-03783-0
28. Salminen A. Activation of immunosuppressive network in the aging process. *Ageing Res Rev*. 2020;57:100998. doi:10.1016/j.arr.2019.100998
29. Darwish M, El-Tamawy MS, Mahmoud A, et al. The impact of physical therapy intervention of dysphagia on preventing pneumonia in acute stroke patients: a randomized controlled trial. *Physiother Res Int*. 2024;29(3):e2108. doi:10.1002/pri.2108
30. Müller ML, Peglau L, Moon LDF, et al. Neurotrophin-3 attenuates human peripheral blood T cell and monocyte activation status and cytokine production post stroke. *Exp Neurol*. 2022;347:113901. doi:10.1016/j.expneurol.2021.113901

31. Tuz AA, Ghosh S, Karsch L, et al. Stroke and myocardial infarction induce neutrophil extracellular trap release disrupting lymphoid organ structure and immunoglobulin secretion. *Nat Cardiovasc Res*. 2024;3(5):525–540. doi:10.1038/s44161-024-00462-8
32. Koupenova M, Corkrey HA, Vitseva O, et al. The role of platelets in mediating a response to human influenza infection. *Nat Commun*. 2019;10 (1):1780. doi:10.1038/s41467-019-09607-x
33. Clark SR, Ma AC, Tavener SA, et al. Platelet TLR4 activates neutrophil extracellular traps to ensnare bacteria in septic blood. *Nat Med*. 2007;13 (4):463–469. doi:10.1038/nm1565
34. Urra X, Cervera A, Villamor N, et al. Harms and benefits of lymphocyte subpopulations in patients with acute stroke. *Neuroscience*. 2009;158 (3):1174–1183. doi:10.1016/j.neuroscience.2008.06.014
35. Xu D, Lu W. Defensins: a double-edged sword in host immunity. *Front Immunol*. 2020;11:764. doi:10.3389/fimmu.2020.00764
36. Wang C, Jiang X, Wu D, et al. GNRI, PLR and stroke-associated pneumonia: from association to development of a web-based dynamic nomogram. *Clin Interv Aging*. 2023;18:1893–1904. doi:10.2147/CIA.S433388
37. Li W, He C. Association of platelet-to-lymphocyte ratio with stroke-associated pneumonia in acute ischemic stroke. *J Healthc Eng*. 2022;2022:1033332. doi:10.1155/2022/1033332