Novel Machine Learning Models to Predict Pneumonia Events in Supratentorial Intracerebral Hemorrhage Populations: An Analysis of the Risa-MIS-ICH Study

Yan Zheng
First Affiliated Hospital of Fujian Medical University
Yuan-xiang Lin
First Affiliated Hospital of Fujian Medical University
Qiu He
First Affiliated Hospital of Fujian Medical University
Chun-wang Li
First Affiliated Hospital of Fujian Medical University
Wei Huang
First Affiliated Hospital of Fujian Medical University
Zhuyu Gao
First Affiliated Hospital of Fujian Medical University
Geng-zhao Ye
First Affiliated Hospital of Fujian Medical University
Ren-long Chen
First Affiliated Hospital of Fujian Medical University
Lve-ming Cai
First Affiliated Hospital of Fujian Medical University
Ming-pei Zhao
First Affiliated Hospital of Fujian Medical University
Ling-yun Zhuo
First Affiliated Hospital of Fujian Medical University
Hao-jie Wang
First Affiliated Hospital of Fujian Medical University
Ze-feng Xie
Anxi County Hospital
Ke Ma
First Affiliated Hospital of Fujian Medical University
Wen-hua Fang
First Affiliated Hospital of Fujian Medical University

Deng-liang Wang
First Affiliated Hospital of Fujian Medical University
Jian-cai Chen
Anxi County Hospital
De-zhi Kang
First Affiliated Hospital of Fujian Medical University
Fu-xin Lin (lfxstudy@126.com)
First Affiliated Hospital of Fujian Medical University

Research Article

Keywords: intracerebral hemorrhage, machine learning, predict, stroke-associated pneumonia

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1542765/v1

Abstract

Background: Stroke-associated pneumonia (SAP) contributes to high mortality rates in spontaneous intracerebral hemorrhage (sICH) populations. The accurate prediction and early intervention of SAP are associated with prognosis. Although various predictive scoring systems have been previously developed, none are widely accepted. We aimed to derive and validate novel supervised machine learning (ML) models to predict SAP events in supratentorial sICH populations.

Methods: In this work, the data of eligible supratentorial sICH individuals were extracted from the database of the *Risa-MIS-ICH* study, and the participants were split into training, internal validation, and external validation datasets. The primary outcome was SAP during hospitalization. Univariate and multivariate analyses were used for variable filtrations, and logistic regression (LR), Gaussian naïve Bayes (GNB), random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM), extreme gradient boosting (XGB), and ensemble soft voting model (ESVM) were adopted for ML model derivations. The metrics of accuracy, sensitivity, specificity, and area under the curve (AUC) were adopted to evaluate the predictive value of each model with internal/cross-/external validations.

Results: After screening 909 individuals with sICH, a total of 468 were included in this work. Six independent variables [nasogastric feeding, airway support, unconscious onset, surgery of external ventricular drainage (EVD), sICH volume, and intensive care unit (ICU) stay] for SAP were identified and selected for seven ML prediction model derivations and validations. The internal and cross-validations revealed the superior and robust performance of the GNB model with the highest AUC value (0.861, *95% CI*: 0.793-0.930), while the LR model had the highest AUC value (0.867, *95% CI*: 0.812-0.923) in external validation. The ESVM method combining the other six methods had moderate but robust abilities in both cross- and external validations and achieved an AUC of 0.843 (*95% CI*: 0.784, 0.902) in the external validation.

Conclusion: The ML models could effectively predict SAP events in sICH populations, and our novel ensemble models demonstrated reliable robust performance outcomes despite the populational and algorithmic differences.

Registration: URL: https://www.clinicaltrials.gov. Unique Identifier: NCT03862729

Introduction

Stroke-associated pneumonia (SAP) is the most common infectious complication in spontaneous intracerebral hemorrhage (sICH) individuals, with an estimated incidence of 15–25% in overall stroke populations [1–3]. SAP is usually adversely associated with increased mortality, prolonged hospital stays, and poor prognosis [3–5]. The current large phase III clinical trials have not found the benefits of routine antibiotic prevention for general stroke individuals [6, 7]. Therefore, the accurate prediction and early intervention of SAP might contribute to improving the prognosis. Thus, a reliable model is needed for

predicting and monitoring potential SAP, so exact prophylactic interventions or therapeutic antibiotics can be tailored in a timely manner.

In recent decades, a few studies have indicated several independent risk factors for SAP, including older age [5, 8–13], male sex [8, 9, 13, 14], severe stroke [4, 5, 8–16], intubation [4, 15], nasogastric feeding or dysphagia [4, 8, 16], and deeper location and larger volume of sICH [4, 11, 15]. Some of these variables were included in several predictive scoring systems for SAP risk stratifications, such as the A^2DS^2 and PNA scores in Germany [9, 12], the AIS/ICH-APS scores in China [10, 11], and the ISAN score in the UK [13]. However, most scoring systems are designed for acute ischemic stroke (AIS) populations [9, 10, 12, 13], and none of the SAP prediction scoring systems is widely accepted in routine clinical practice.

Machine learning (ML) is the key approach to solving artificial intelligence (AI) problems. Compared with traditional scoring systems, ML models have shown better performance in predicting disease occurrence and prognosis by adapting known cases and invoking their experience [17–19]. The advantages of intelligent algorithms and extreme data usage allow ML to manage nonlinear, high-dimensional, and even undiscovered correlations in datasets. The implementation of ML algorithms could be beneficial in identifying significant predictors for the automation of cumbersome clinical assessments [19, 20]. In this work, we aim to derive and validate novel supervised ML models to predict SAP events in supratentorial sICH populations and expect to develop a superior and automatic tool for clinical practice.

Methods

Study Design and Participants

The data for this analysis were obtained from the retrospective database of the *Risk Stratification and Minimally Invasive Surgery in Acute Intracerebral Hemorrhage Patients* (*Risa-MIS-ICH*) study (ClinicalTrials.gov Identifier: NCT03862729), which was a multicenter ambispective cohort study. Two centers were involved in the retrospective cohort for this work, including First Affiliated Hospital, Fujian Medical University (FAHFMU, Fuzhou, Fujian), and Anxi County Hospital (ACH, Quanzhou, Fujian). The study protocol followed the principles of the Declaration of Helsinki and was approved by the ethics committee of FAHFMU (GN: MRCTA, ECFAH of FUM [2018]082) and documented in each center. No informed consent was required for the retrospective cohort. This work was reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement [21].

A total of 729 individuals were diagnosed with supratentorial sICH and received therapies at FAHFMU (from January 2015 to January 2021) and were included in the FAHFMU subcohort for the variable filtrations and model derivations/validations. An external subcohort with 180 participants from ACH (from June 2019 to January 2021) was introduced into this work for external validation.

The inclusion and exclusion criteria of the participants are shown in Table 1, and the screening process is presented in Fig. 1A.

Table 1
Eligibility criteria for retrospective cohort enrollment in the Risa-MIS-ICH study and current work

Inclusion criteria:
(1) Diagnosed with spontaneous sICH with CT/CTA scan confirmation, and the interval time from onset to recorded CT/CTA scan ≤ 48 hours (2) GCS score > 5 and no cerebral herniation at admission (3) Onset age ≥ 18 years
Exclusion criteria:
(1) With any intracranial etiology of supratentorial hemorrhage of AVM, arterial aneurysm, hemorrhagic cerebral tumor stroke, hemorrhagic infarction, coagulation disorders, or any other potential organic lesions indicating nonspontaneous sICH (2) Occurrence of infratentorial hemorrhage (3) Evidence of pregnancy, or prestroke life expectancy < 3 months.
Extra criteria for SAP prediction model derivations/validations in this work:
(1) Interval time from onset to admission ≤ 24 hours (2) Hospital stays ≥ 48 hours (3) Receiving no mechanical ventilation or ventilation time ≤ 24 hours before SAP events (4) Key data loss about SAP in laboratory, imaging, or other important clinical information
sICH: supratentorial intracerebral hemorrhage, CT: computed tomography, CTA: CT angiography, GCS: Glasgow Coma Scale, AVM: arteriovenous malformation, SAP: stroke-associated pneumonia.

Variable Extractions and Primary Outcomes

Relevant information about participants was retrieved from the electronic medical record (EMR) systems from each neurological research center. The electronic data capture (EDC, RealData Corporation, Ningbo, Zhejiang, P.R. China) system was employed for database establishment and data collection. The trained professional clinical research coordinators (CRCs) were commissioned for data entry and follow-up. The *Risa-MIS-ICH* database included 665 variables and involved information on demographics, prestroke comorbidities, onset details, imaging features, laboratory results, complications during hospitalizations, interventions, discharge status, and follow-up information. The collation of the database was performed by professional statisticians, and data analysis was carried out after passing the third-party quality control.

The primary outcome of the current analysis was the occurrence of SAP events during hospitalization, and SAP was defined as pneumonia not incubating during hospital admission and occurring ≥ 48 hours

after admission in acute stroke populations. Referring to the diagnostic criteria for hospital-acquired pneumonia (HAP), the diagnostic criteria for SAP were as follows [22, 23]: the presence of a new or progressive infiltrate in chest X-ray or computed tomography (CT) scan, plus at least two of the following clinical manifestations: (1) fever (T > 38°C) or hypothermia (T < 36°C), (2) leukocytosis [white blood cell (WBC) count > 10×10^9/L] or leukopenia (WBC count < 4×10^9/L), and (3) nursing-recorded purulent airway secretion. Ventilator-acquired pneumonia (VAP), defined as a pneumonia event after ventilation time > 24 hours, was excluded from this work.

Statistical Analysis and Variable Filtration

All statistical analyses were performed using SPSS software (version 22.0, IBM Corporation, Armonk, NY, USA) and Python (version 3.8.3, Anaconda Distribution, Austin, TX, USA). The current work mainly used the development environment of Jupyter Notebook (version 6.0.3) and invoked the key packaged libraries of NumPy (version 1.18.5), Pandas (version 1.1.5), Scikit-learn (version 0.24.2), SciPy (version 1.5.0), Matplotlib (version 3.4.3), and Lifelines (0.26.4). The continuous variables and categorical variables are presented as the mean and standard deviation (SD) or median and interquartile range (IQR) and quantities and percentages.

The screening of variables was performed in the FAHFMU subcohort. As shown in Fig. 1B, the study variables were initially screened by univariate analyses. The independent sample Student's *t test* was used for normally distributed data, the *Mann–Whitney U* test was used for nonnormally distributed data, and the chi-square test or Fisher's exact test was used for categorical data. All tests in this work were two-sided, and $P < 0.05$ was considered statistically significant. To prevent overfitting, least absolute shrinkage and selection operator (LASSO) regression was used in multivariate analysis and further performed after univariate analyses. Each continuous variable was standardized before performing LASSO regression to improve generalizability. LASSO regression selects the optimal penalty value *via* the internally installed *k*-fold cross-validation module ($k = 3$) and recursively removes the least important variables by vanishing coefficients. Through the above steps, the independent significant variables had nonzero coefficients in LASSO regression and were selected as candidate variables for ML model derivations.

Survival analysis was additionally performed in this work, in which all-cause death after stroke onset was defined as the observed indicator. The survival time was defined as the time interval from stroke onset to all-cause death or follow-up. The survival curves were plotted using the Kaplan–Meier method, and survival rates were compared using the log-rank test.

Model Derivations and Validations

The flow diagram of the model derivations and validations is presented in Fig. 1C. The FAHFMU subcohort was randomly split into the training and validation datasets (7:3), which were used for the model derivations and the internal validation, respectively. The model derivations were performed on the candidate variables by six common basic ML algorithms and one additional ensemble model, including logistic regression (LR), Gaussian naïve Bayes (GNB), random forest (RF), K-nearest neighbor (KNN),

support vector machine (SVM), extreme gradient boosting (XGB), and ensemble soft voting model (ESVM). None of these models was uncertain about demonstrating the optimal performance beforehand. In the training process, six basic ML algorithms were independently fitted with the candidate variables and virtual SAP classifications from the training dataset, and model hyperparameters were optimized with the grid-search algorithm to promote model performance. In detail, the grid-search algorithms tune optimal parameters by internally evaluating model performance repeatedly *via* the nested *k*-fold cross-validation module (*k* = 3 in this work). Before to the above steps, ML prediction models with different characteristics were generated, and these processes were termed supervised ML. To improve the robustness of ML models, the additional ESVM was derived incorporating the aforementioned six algorithms. The ESVM is a simply voting system on the weighted classified outputs of the six basic algorithms, and these processes were termed soft voting.

After model derivations, the validation dataset was automatically inputted into the seven models to obtain the predicted classifications in the internal validation. Receiver operating characteristic (ROC) curves were plotted, and the metrics of accuracy, sensitivity, specificity, and area under the curve (AUC) along with 95% confidence intervals (CIs) were adopted to evaluate the predictive value of each model with validations. Further supplementary internal evaluation with advanced robustness was performed with *n*-repeated *k*-fold cross-validation (*n* = 3 and *k* = 5 in this work). This method repartitions the FAHFMU subcohort into *k* nonoverlapping folds, where the *k*-1 folds are used for the model derivations and the other fold is used for validation. After *n* repetitions, *n*×*k* combinations are finally generated for robust validation [24, 25].

Furthermore, this work also introduced the external subcohort, which was not involved in variable filtrations and model derivations. In this process, the entire FAHFMU subcohort was considered the training dataset to retrain the prediction models, and the external subcohort was introduced as the exclusive validation dataset. The technical avenue of training and evaluating the models remained the same as above.

Results

Participants and Characteristics

From January 2015 to January 2021, a total of 909 participants were included in the retrospective cohort of the *Risa-MIS-ICH* study, and 441 of these individuals were excluded due to ventilation > 24 hours, ineligible time window, or incomplete data. Finally, four hundred and sixty-eight individuals (n_{FAHFMU} = 324, n_{ACH} = 144) were retained for this work. The overall average age was 60.44 (± 12.51) years, and 308 (65.8%) of the individuals were male. SAP events during hospitalizations occurred in 135 (28.8%) [n_{FAHFMU} = 97 (29.9%), n_{ACH} = 38 (26.4%)] individuals. The demographic characteristics, clinical manifestations, imaging features, laboratory tests, and prognostic indicators in the FAHFMU and external subcohorts are summarized in Tables 2 and 3, respectively. Differences in the analyzed variables between the two centers are shown in Supplemental Table 1.

Table 2
Baseline characteristics

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
Age (years)	58.6 (± 11.8)	60.0 (± 12.6)	0.370	62.7 (± 12.7)	66.0 ± (13.5)	0.182
Sex						
Male (n)	155 (68.3%)	69 (71.1%)	0.694	59 (55.7%)	25 (65.8%)	0.339
Female (n)	72 (31.7%)	28 (28.9%)		47 (44.3%)	13 (34.2%)	
Prestroke History						
Hypertension (n)	163 (71.8%)	74 (76.3%)	0.416	65 (61.3%)	27 (71.7%)	0.329
Diabetes Mellitus (n)	29 (12.8%)	13 (13.4%)	1.000	4 (3.8%)	3 (7.9%)	0.381
Heart Disease (n)	8 (3.5%)	4 (4.1%)	1.000	2 (1.9%)	2 (5.3%)	0.284
Smoking (n)	59 (26.0%)	24 (24.7%)	0.890	1 (0.9%)	0	1.000
Alcohol Abuse (n)	59 (26.0%)	23 (23.7%)	0.679	1 (0.9%)	0	1.000
Previous Surgery (n)	48 (21.1%)	19 (19.6%)	0.768	2 (1.9%)	4 (10.5%)	**0.042**
Onset Form						
Neurological Dysfunction (n)	201 (88.5%)	72 (74.2%)	**0.002**	91 (85.8%)	31 (81.6%)	0.600
Unconsciousness (n)	54 (23.8%)	71 (73.2%)	**< 0.001**	27 (25.5%)	27 (71.1%)	**< 0.001**
Epileptic Attack (n)	4 (1.8%)	4 (4.1%)	0.246	2 (1.9%)	0	1.000
Headache (n)	71 (31.3%)	24 (24.7%)	0.287	91 (85.8%)	21 (55.3%)	**< 0.001**
Others (n)	93 (41.0%)	39 (40.2%)	0.903	94 (88.7%)	29 (76.3%)	0.105
Interval Time from Onset to Admission (h)	12.0 (7.0, 24.0)	10.0 (6.5, 16.0)	**0.022**	3.0 (2.0, 8.3)	3.0 (2.0, 4.5)	0.103

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
Admission Examination						
Temperature (°C)	36.5 (36.5, 36.8)	36.7(36.5, 36.9)	0.115	36.6 (36.5, 36.8)	36.6 (36.5, 36.7)	0.667
Heart Rate (min^{-1})	77 (± 14)	83 (± 17)	**0.002**	81 (± 12)	84 (± 14)	0.237
Respiratory Rate (min^{-1})	20(19, 20)	20(19, 21)	**0.008**	20 (20, 20)	20 (20, 20)	0.998
Systolic BP (mmHg)	158 (± 24)	162 (± 25)	0.145	170 (± 24)	174 (± 27)	0.473
Dilated BP (mmHg)	93 (± 15)	92 (± 14)	0.610	100 (± 15)	101.8 (± 16)	0.453
Admission GCS Score						
15 (n)	106 (46.7%)	12 (12.4%)	**< 0.001**	80 (75.5%)	10 (26.3%)	**< 0.001**
13–14 (n)	77 (33.9%)	33 (34.0%)		8 (7.5%)	5 (13.2%)	
9–12 (n)	31 (13.7%)	19 (19.6%)		14 (13.2%)	15 (39.5%)	
5–8 (n)	13 (5.7%)	33 (34.0%)		4 (3.8%)	8 (21.1%)	
Hospital Costs (thousand CNY)*	17.0 (12.5, 25.8)	49.7 (34.4, 91.0)	**< 0.001**	7.7 (6.5, 10.8)	25.1 (14.6, 35.7)	**< 0.001**
Hospital Stays (d)*	15 (11, 20)	17 (13, 24)	**0.003**	14 (12, 15)	23 (15, 29)	**< 0.001**
Discharge Status*						
Home/Nursing or Rehabilitation (n)*	96 (42.3%)	46 (47.6%)	0.463	97 (91.5%)	29 (76.3%)	**0.022**
Care Withdrawal or Hospital Death (n)*	131 (57.7%)	51 (52.6%)		9 (8.5%)	9 (23.7%)	
Mortality (since onset)*						
Survival ≥ 1 year(n)*	168 (74.0%)	63 (64.9%)	**0.009**	77 (72.6%)	20 (52.6%)	**0.013**

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
3 Months – 1 year (n)*	4 (1.8%)	6 (6.2%)		2 (1.9%)	2 (5.3%)	
< 3 Months (n)*	7 (3.1%)	10 (10.3%)		1 (0.9%)	4 (10.5%)	
Loss of Follow-up (n)*	48 (21.1%)	18 (18.6%)		26 (24.5%)	12 (31.6%)	

*These prognostic variables were not included in further multivariate analysis and model derivations/validations.

BP: blood pressure, GCS: Glasgow Coma Scale, CNY: Chinese yuan.

Table 3
Variables of laboratory results, imaging features, and early clinical interventions

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
RBC (10^{12} L^{-1})	4.66 (4.30, 4.94)	4.59 (4.15, 4.87)	0.097	4.66 (4.29, 5.08)	4.64 (4.30, 5.11)	0.928
Hemoglobin ($g \cdot L^{-1}$)	142.2 (± 14.2)	140.2 (± 15.3)	0.278	139.9 (± 17.2)	137.9 (± 20.2)	0.553
Hematocrit	0.41 (± 0.04)	0.41 (± 0.04)	0.681	0.42 (± 0.05)	0.41 (± 0.05)	0.335
WBC (10^9 L^{-1})*	8.52 (6.61, 10.64)	10.17 (7.54, 13.01)	**< 0.001**	8.27 (6.62, 10.83)	9.95 (7.77, 12.28)	**0.014**
Neutrophil (10^9 L^{-1})	6.46 (4.42, 8.72)	8.46 (5.49, 11.61)	**< 0.001**	5.86 (4.41, 8.45)	7.49 (5.03, 10.38)	**0.016**
Lymphocyte (10^9 L^{-1})	1.29 (0.86, 1.66)	1.04 (0.7, 1.39)	**0.001**	1.37 (0.99, 1.82)	1.46 (0.99, 1.90)	0.724
Platelet (10^9 L^{-1})	217.4 (± 62.3)	214.8 (± 63.3)	0.897	235.1 (± 62.4)	221.2 (± 55.7)	0.227
PT (s)	11.1 (10.8, 11.7)	11.1 (10.6, 11.9)	0.925	11.3 (10.9, 11.8)	11.4 (10.9, 12.2)	0.307
PT-INR	0.97 (0.94, 1.02)	0.97 (0.93, 1.04)	0.554	0.98 (0.94, 1.03)	0.99 (0.94, 1.07)	0.294
APTT (s)	25.0 (22.2, 27.9)	24.1(21.8, 27.2)	0.200	25.3 (23.9, 27.1)	24.8 (23.2, 26.8)	0.385
Fibrinogen ($g \cdot L^{-1}$)	2.64 (2.23, 3.04)	2.69(2.30, 3.13)	0.677	2.62 (2.20, 3.12)	2.68 (2.35, 3.18)	0.607
Serum Creatinine ($\mu mol \cdot L^{-1}$)	67.0(54.0, 78.3)	66.0 (54.7, 78.2)	0.769	66.0 (57.0, 82.0)	71.5 (58.8, 95.0)	0.098
Serum Urea Nitrogen ($mmol \cdot L^{-1}$)	5.02 (4.13, 5.94)	5.15(4.27, 6.59)	0.259	4.85 (4.00, 5.83)	5.10 (4.28, 6.85)	0.276

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
Serum Sodium ($mmol \cdot L^{-1}$)	139.5 (± 3.9)	139.9 (± 4.6)	0.486	138.7 (± 3.5)	138.1 (± 3.1)	0.386
Serum Potassium ($mmol \cdot L^{-1}$)	3.80 (± 0.42)	3.84 (± 0.47)	0.474	3.88 (± 0.53)	3.92 (± 0.61)	0.723
Serum Calcium ($mmol \cdot L^{-1}$)	2.28 (± 0.54)	2.20 (± 0.13)	0.158	2.36 (± 0.12)	2.36 (± 0.15)	0.802
Serum Chloride ($mmol \cdot L^{-1}$)	102.0 (99.0, 105.0)	102.6 (99.0, 105.0)	0.743	100.6 (97.8, 102.5)	99.4 (96.3, 101.4)	0.058
sICH Volume (cc)	8.7 (3.9, 17.2)	22.5 (9.4, 37.9)	**< 0.001**	6.8 (3.5, 13.4)	21.7 (6.3, 40.4)	**< 0.001**
Lobar Involvement (n)*	38 (16.7%)	25 (25.8%)	0.067	23 (21.7%)	12 (31.6%)	0.271
Frontal Lobe (n)	17 (7.5%)	14 (14.4%)	0.063	8 (7.5%)	5 (13.2%)	0.328
Parietal Lobe (n)	15 (6.6%)	13 (13.4%)	0.054	10 (9.4%)	4 (10.5%)	1.000
Temporal Lobe (n)	17 (7.5%)	14 (14.4%)	0.063	10 (9.4%)	9 (23.47%)	**0.047**
Occipital Lobe (n)	7 (3.1%)	3 (3.1%)	1.000	5 (4.7%)	2 (5.3%)	1.000
Deep Involvement (n)*	204 (89.9%)	87 (89.7%)	1.000	87 (82.1%)	31 (81.6%)	1.000
Basal Ganglia (n)	174 (76.7%)	74 (76.3%)	1.000	66 (62.3%)	29 (76.3%)	0.162
Thalamus (n)	56 (24.7%)	33 (34.0%)	0.103	33 (31.1%)	11 (28.9%)	0.841
Corona Radiata (n)	5 (2.2%)	4 (4.1%)	0.552	6 (5.7%)	6 (15.8%)	0.082
Insular Lobe (n)	4 (1.8%)	1 (1.0%)	1.000	9 (8.5%)	6 (15.8%)	0.223
Intraventricular Involvement (n)*	60 (26.4%)	47 (48.5%)	**< 0.001**	37 (34.9%)	15 (39.5%)	0.695

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	
Unilateral Ventricle (n)	26 (11.5%)	13 (13.4%)	**< 0.001**	21 (19.8%)	7 (18.4%)	0.227
Bilateral Ventricles (n)	33 (14.5%)	33 (34.0%)		15 (14.2%)	8 (21.1%)	
Third Ventricle (n)	29 (12.8%)	26 (26.8%)	**0.003**	17 (16.0%)	10 (26.3%)	0.224
Fourth Ventricle (n)	24 (10.6%)	22 (22.7%)	**0.006**	14 (13.2%)	7 (18.4%)	0.593
Subarachnoid Involvement (n)	7 (3.1%)	8 (8.2%)	**0.050**	3 (2.8%)	1 (2.6%)	1.000
ICU Stay (n)	14 (6.2%)	39 (40.2%)	**< 0.001**	0	8 (21.1%)	**< 0.001**
Nasogastric Feeding (n)	59 (26.0%)	84 (86.6%)	**< 0.001**	11 (10.4%)	24 (63.2)	**< 0.001**
Airway Support						
None (n)	215 (94.7%)	48 (49.5%)	**< 0.001**	105 (99.1%)	30 (78.9%)	**< 0.001**
Endotracheal Intubation ≤ 24 hours or Naso-/Oropharyngeal Airway (n)	2 (0.9%)	13 (13.4%)		0	4 (10.5%)	
Endotracheal Intubation > 24 hours or Tracheotomy (n)	10 (4.4%)	36 (37.1%)		1 (0.9%)	4 (10.5%)	
Surgery*	18 (7.9%)	50 (51.5%)	**< 0.001**	14 (13.2%)	22 (57.9%)	**< 0.001**
Only sICH Evacuation (n)	11 (4.8%)	20 (20.6%)	**< 0.001**	0	4 (10.5%)	**0.004**
Only Endoscopic sICH Evacuation (n)	1 (0.4%)	1 (1.0%)	0.510	0	0	-
Only sICH Catheter Evacuation (n)	0	2 (2.1%)	0.089	9 (8.5%)	7 (18.4%)	0.089
Only EVD Approach (n)	4 (1.8%)	15 (15.5%)	**< 0.001**	3 (2.8%)	9 (23.7%)	**< 0.001**
Ensemble Approaches (n)	2 (0.9%)	12 (12.4%)	**< 0.001**	2 (1.9%)	2 (5.3%)	0.573

Variables	FAHFUM Subcohort		P value	External Subcohort		P value
	Without SAP (n = 227)	With SAP (n = 97)		Without SAP (n = 106)	With SAP (n = 38)	

*These general variables were not included in further multivariate analysis and model derivations/validations.

RBC: red blood cell; WBC: white blood cell; PT: prothrombin time; INR: international normalized ratio; APTT: activated partial thromboplastin time; sICH: spontaneous intracerebral hemorrhage; ICU: intensive care unit; EVD: external ventricular drainage.

Table 4
Performance metrics of the ML models in the FAHFUM validation dataset and external subcohort

	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
(A) Internal Validation				
LR	0.838 (0.765,0.911)	0.827 (0.752,0.901)	0.615 (0.519,0.712)	0.903 (0.844,0.961)
GNB	0.861 (0.793,0.930)	0.816 (0.740,0.893)	0.615 (0.519,0.712)	0.889 (0.827,0.951)
RF	0.837 (0.763,0.910)	0.816 (0.740,0.893)	0.462 (0.363,0.560)	0.944 (0.899,0.990)
KNN	0.807 (0.729,0.885)	0.786 (0.704,0.867)	0.500 (0.401,0.599)	0.889 (0.827,0.951)
SVM	0.770 (0.687,0.854)	0.786 (0.704,0.867)	0.500 (0.401,0.599)	0.889 (0.827,0.951)
XGB	0.839 (0.766,0.912)	0.827 (0.752,0.901)	0.692 (0.601,0.784)	0.875 (0.810,0.940)
ESVM	0.830 (0.756,0.904)	0.837 (0.764,0.910)	0.615 (0.519,0.712)	0.917 (0.862,0.971)
(B) External Validation				
LR	0.867 (0.812,0.923)	0.812 (0.749,0.876)	0.447 (0.366,0.529)	0.943 (0.906,0.981)
GNB	0.856 (0.798,0.913)	0.833 (0.772,0.894)	0.553 (0.471,0.634)	0.934 (0.893,0.975)
RF	0.844 (0.784,0.903)	0.806 (0.741,0.870)	0.368 (0.290,0.447)	0.962 (0.931,0.993)
KNN	0.734 (0.662,0.806)	0.778 (0.710,0.846)	0.395 (0.315,0.475)	0.915 (0.870,0.961)
SVM	0.730 (0.658,0.803)	0.778 (0.710,0.846)	0.395 (0.315,0.475)	0.915 (0.870,0.961)
XGB	0.856 (0.799,0.913)	0.792 (0.725,0.858)	0.421 (0.340,0.502)	0.925 (0.881,0.968)
ESVM	0.843 (0.784,0.902)	0.812 (0.749,0.876)	0.447 (0.366,0.529)	0.943 (0.906,0.981)

ML: machine learning, AUC: area under the curve, LR: logistic regression, GNB: Gaussian naïve Bayes, RF: random forest, KNN: K-nearest neighbor, SVM: support vector machine, XGB: extreme gradient boosting, ESVM: ensemble soft voting model.

Variable Filtration and Importance

Seventy of 665 variables were retained for subsequent analyses. Univariate analysis showed that 25 variables were potential predictive factors for SAP (Tables 2 and 3). The retained variables of nasogastric feeding, airway support, unconscious onset, surgery of external ventricular drainage (EVD), sICH volume (estimated with the ABC/2 formula in imaging), and intensive care unit (ICU) stay were considered independent predictors for SAP in LASSO regression. These candidate variables with ranked coefficients are shown in Fig. 2. The evolution of tuning parameter optimization in LASSO regression is shown in Fig. 3.

Model Performance

Internal validation was performed with the quantified metrics and visualized ROC curves, as shown in Table 4A and Fig. 4A. Among the seven models of LR, GNB, RF, KNN, SVM, XGB, and ESVM, GNB demonstrated the optimal efficiency to predict SAP with the highest AUC value (0.861, *95% CI*: 0.793–0.930), while the ESVM presented the highest accuracy (0.837, *95% CI*: 0.764–0.910) and specificity (0.917, *95% CI*: 0.862–0.971). The XGB was the most sensitive, with the highest value (0.692, *95% CI*: 0.601–0.784). The learning curves presented the evolutions of models with different characteristics and are illustrated in Fig. 5.

Three repeated fivefold cross-validation were established, and a total of 15 combinations were generated from three splits and five folds. The AUC values of different models from combinations are summarized and presented as heatmaps in Fig. 6, and all quantified metrics are listed in Supplemental Table 2. In most random states, the ESVM (n = 9) and XGB (n = 8) models remained the optimal models in terms of accuracy and sensitivity, respectively. Unlike the results in internal validation, the LR (n = 6) and RF (n = 10) models most often had the highest AUC and specificity values, respectively, with robustness.

External Validation

The metrics and ROC curves of each model in external validation are shown in Table 4B and Fig. 4B. The LR was superior in AUC value (0.867, *95% CI*: 0.812–0.923) in the external validation. While GNB had the highest accuracy (0.833, *95% CI*: 0.772–0.894) and sensitivity (0.553, *95% CI*: 0.471–0.634), the RF was the most specific (0.962, *95% CI*: 0.931–0.993). There was no single algorithm with dominant ability and robustness in the external validation. It is worth mentioning that the ESVM had moderate but robust abilities and achieved AUC, accuracy, sensitivity, and specificity values of 0.843 (*95% CI*: 0.784–0.902), 0.812 (*95% CI*: 0.749–0.876), 0.447 (*95% CI*: 0.366–0.529), and 0.943 (*95% CI*: 0.906–0.981), respectively, in the external validation.

Outcome and Survival Analysis

In both the FAHFUM and external subcohorts, participants with SAP suffered from significantly higher hospital costs and prolonged hospital stays (both $P < 0.001$). Three hundred sixty-four (77.8%) of all

eligible 468 participants were followed for survival, and 83 (25.3%) of them had experienced SAP during hospitalization. The average survival times were 44.95 ± 2.78 (*95% CI*: 39.50–50.40) and 55.77 ± 1.26 (*95% CI*: 53.30-58.25) in participants with and without SAP, respectively. The median survival times were not available due to death < 50%. Both 3-month (86.9% *vs.* 96.7%) and 1-year (78.3% *vs.* 94.2%) survival rates were lower in participants with SAP than in those without SAP, and the log-rank test showed significant discrimination ($P < 0.001$) of overall survival between the two groups. The Kaplan-Meier curves are plotted in Fig. 7.

Discussion

It is critical to identify individuals at high risk for SAP and to further tailor timely prophylactic interventions or therapeutic antibiotics. However, for now, the early prediction of SAP in sICH populations is challenging due to the lack of widely accepted prediction tools, which are important for modern precision medicine and evidence-based medicine (EBM) in this field. Thus, we aimed to derive more effective and automatic sICH-SAP prediction tools in this work. The novel ML prediction models were derived and validated as an attempt to combine AI medical engineering and clinical practice in this field. The major findings were as follows. (1) The incidence rate of sICH-SAP was close to 30%, and the sICH-SAP events significantly contributed to prolonged hospital stays, increased hospital costs, and higher mortality. (2) Six independent predictors for sICH-SAP were identified — nasogastric feeding, airway support, unconscious onset, surgery of EVD, sICH volume, and ICU stay. (3) ML prediction models were successfully derived and showed better performance metrics than traditional scoring systems from previous studies; the GNB and LR models showed the highest AUC values of 0.861 (*95% CI*: 0.793–0.930) and 0.867 (*95% CI*: 0.812–0.923) on the internal and external validation datasets, respectively. (4) There was no certain single algorithm with dominant ability and robustness in cross- and external validations, while the ESVM was considered averaged in metrics and robust in different populations after multiple validations.

Various predictors for SAP were identified in prior literature [4, 5, 8–16]. This work screened for independent variables for sICH-SAP events by using univariate and multivariate analyses in the FAHFMU subcohort. Nasogastric feeding, airway support, and unconscious onset were identified as strongly associated risk predictors, which overlapped with the results of previous studies [4, 8–16]. Nasogastric feeding and airway support measurement were recognized as SAP predictors, which might bring about secretion disturbances in nasal/oral/tracheal cavities, decreased air filtrations, and even aspiration events [4, 8, 15, 16]. These early interventions were secondary to the manifestation of unconsciousness. Previous studies mainly included the ranked variable of the Glasgow Coma Scale (GCS) score and rarely adopted the onset manifestations [4, 10, 11, 14–16]. In this work, the admission GCS score and unconscious onset were simultaneously introduced into the analyses, and the categorical variable of unconscious onset was independently significant for sICH-SAP. The predictors of sICH volume and ICU stay were also reported in previous studies [4, 11, 15] and contributed the least to predicting SAP in this work. The sICH volume resulted in SAP being a primary factor influencing stroke severity, and ICU stay was a comprehensive intervention secondary to stroke severity and resulted in infectious environments.

These aforementioned predictors are usually uncontrollable for actively preventing SAP in clinical practice. However, there were still novel findings in the subgroup analysis that only the surgery of EVD was a significant independent predictor ($P < 0.001$ in FAHFUM/$P = 0.001$ in external subcohorts) of all surgical approaches in this work, while EVD was only previously reported as a univariate factor for overall infections [4]. On the other hand, the surgery of sICH catheter evacuation did not significantly contribute to SAP events in any univariate analyses (both $P = 0.089$ in FAHFUM/external subcohorts), which was in accordance with the undifferentiated non-neurologic infections in the *MISTIE* III trial [26]. This suggests that we should continuously focus on the stratification of surgical approaches in the prospective cohort of the *Risa-MIS-ICH* study for convincing evidence.

We observed that there was populational heterogeneity from multiple centers, which might result in variations in demographic features, stroke severity, and even baseline laboratory results and further inconsistently significant results in univariate analyses. Interestingly, heterogeneity was not found in these six independent variables, which were effectively predictive in external validation. Therefore, these six independent variables were considered robust in different populations, and robustness evaluations for further ML models became possible in this work.

To date, none of the SAP prediction models have been widely available in clinical practice, and only the ICH-APS score has been developed for sICH populations as a mature SAP prediction model [8–13]. The ICH-APS score also included early indicators, and the AUC value was 0.74 (*95% CI*: 0.72–0.75) on its original validation dataset from the China National Stroke Registry (CNSR). In this work, our optimal ML prediction models achieved higher AUC values of 0.861 (*95% CI*: 0.793–0.930) and 0.867 (*95% CI*: 0.812–0.923) in the internal and external validations, respectively. Our ML prediction models showed greater predictive abilities than the ICH-APS score on their original validation datasets.

Li *et al* developed ML models to predict SAP events in Chinese AIS populations, which presented better performance with the highest AUC value of 0.843 (*95% CI*: 0.803–0.882) than other AIS-SAP prediction scores (0.835 for A^2DS^2, 0.786 for PNA, 0.785 for AIS-APS, and 0.78 for ISAN scores) [27]. According to metrics from the literature and this work [27–30], the ML prediction models for SAP showed better performance metrics than traditional scoring systems in both sICH and AIS populations. However, due to incomplete variable collections, horizontal comparisons of different prediction models on the same validation dataset were not possible. Despite the defects, the prediction models usually performed better in internal validation than in external validation due to the intrinsic consistency of original datasets and populational heterogeneity, and the comparisons on their respective original validation datasets usually explained the significance [31].

In this work, we used six separate algorithms and an ensemble model for SAP predictions, each of which has been validated and compared to identify the optimal prediction tool. For the six basic algorithms, LR, GNB, RF, and XGB generally showed better performance metrics than KNN or SVM in the internal/cross-/external validations, and LR and GNB required less training time than the other algorithms. It is worth mentioning that there was no certain model with the dominant ability and

robustness among internal/cross-/external validations, and the indeterminacy restricted the previous model selection and implementation in clinical practice. Therefore, a general and robust model is required for stable predictive ability. The traditional research mainly focused on the mutually separated algorithms, and only the optimal algorithm was chosen as the option, while ensemble ML models were reported as successful classifiers with greater performance outcomes in the literature [29, 30]. In the real world, there would be no fault-tolerant chance of model selections due to ethical considerations. The predictive ability of one single algorithm was uncertain due to the inconsistent ML algorithmic performance outcomes among the internal/cross-/external validations. Thus, we additionally derived ESVM based on a soft voting system incorporating six basic ML algorithms, which was moderate but surprisingly robust in each metric. Notwithstanding that the occupied machine sources of the ESVM equals the summation of the six basic algorithms, this disadvantage could be ignored by timed training and then *pro re nata* invoking.

A practical ML prediction model requires high accuracy and automation in the real world, which might represent the main directions of cross-nested algorithm deepening and AI medical engineering development in the coming decades. The synthetic minority oversampling technique (SMOTE) and principal component analysis (PCA), which aimed to adjust imbalanced classifications and reduce data dimensions to reduce overfitting and the training time, respectively, were tried but abandoned for reasons in this work [32, 33]. However, our failed attempts and other undiscovered advanced AI techniques could be employed in other AI medical research and clinical practice in the near future. By dynamically evaluating the keyed-in clinical manifestations, the resulting values from the laboratory information system (LIS) and the captured data from the picture archiving and communication system (PACS), the internally installed sophisticated algorithms in the EMR system would ceaselessly learn and then calculate the prediction for high-risk individuals in the prospect. The attempts and prospects in this work might be considered as the progression of automatic clinical evaluations and AI-assisted decision-making.

We have strengths that deserve comments. An external subcohort and multiple forms of validation were introduced in this work. Therefore, there were populational and algorithmic robustness of convincing results. Based on the aforementioned circumstances, we derived novel ensemble models for generalizability, which showed moderate but robust predictive abilities in different populations and were fit for real-world practice. However, there are limitations that should be acknowledged in this work. First, the observational retrospective design might introduce unmanageable bias. Uncontrollable baseline characteristics in the observational study might confound SAP risks and further model derivations/validations. Second, some important variables were missing due to the retrospective collection of data in this work. The National Institute of Health Stroke Scale (NIHSS) score, uniform CT scan parameters, scanning timing, and other unrecorded details were unreachable in the retrospective cohort of the *Risa-MIS-ICH* study and resulted in the inability to perform horizontal comparisons with external models in this work. Third, there are defects of the deep analyses for SAP. The subgroup analyses on pneumonia severity, radiological features, or pathogenic agents were all absent. A simple

overall SAP analysis might be rather rough for complex and heterogenetic pulmonary infections. Future studies on our ongoing prospective cohort might resolve the aforementioned problems.

To the best of our knowledge, this was the first reported attempt of ML prediction model for sICH-SAP. The authors not only aimed to derive superior statistical models for SAP prediction but also attempted to combine AI medical engineering and clinical practice in this field. We truly anticipate that this technique will be developed as an effective and automatic tool for predicting sICH-SAP in the near future.

Conclusion

In this work, the authors derived SAP prediction models with ML algorithms in supratentorial sICH populations from multiple centers and performed multiple validations for effective and robust confirmations. The ensemble model was novelly employed in this work and showed robust performance outcomes in different populations.

Abbreviations

SAP: stroke-associated pneumonia, sICH: spontaneous intracerebral hemorrhage, AIS: acute ischemic stroke, ML: machine learning, AI: artificial intelligence, TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis, EMR: electronic medical record, EDC: electronic data capture, CRC: clinical research coordinator, HAP: hospital-acquired pneumonia, CT: computed tomography, WBC, white blood cell, VAP: ventilator-acquired pneumonia, SD: standard deviation, IQR: interquartile range, LASSO: least absolute shrinkage and selection operator, LR: logistic regression, GNB: Gaussian naïve Bayes, RF: random forest, KNN: K-nearest neighbor, SVM: support vector machine, XGB: extreme gradient boosting, ESVM: ensemble soft voting model, ROC: receiver operating characteristic, AUC: area under the curve, CI: confidence interval, EVD: external ventricular drainage, ICU: intensive care unit, EBM: evidence-based medicine, GCS: Glasgow Coma Scale, CNSR: China National Stroke Registry, SMOTE: synthetic minority oversampling technique, PCA: principal component analysis, LIS: laboratory information system, PACS: picture archiving and communication system, NIHSS: National Institute of Health Stroke Scale.

Declarations

Ethics Approval and Consent to Participate

The study protocol followed the principles of the Declaration of Helsinki and was approved by the ethics committee of FAHFMU (GN: MRCTA, ECFAH of FUM [2018]082) and documented in each center. No informed consent was required for the retrospective cohort.

Consent for Publication

Not applicable.

Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors declare no financial or other competing interests.

Source of Funding

This work was supported by the Popularization of Appropriate Intervention Technology for the Stroke High Risk Group in China from the Stroke Prevention and Treatment Project of the National Health Commission (GN-2018R0002), the Technology Platform Construction Project of Fujian Province from the Fujian Institute for Brain Disorders and Brain Science (2021Y2001), the Technology Platform Construction Project of Fujian Province from the Fujian Provincial Clinical Research Center for Neurological Diseases (2020Y2003), and the Provincial High Level Neuromedical Center Construction Fund of Fujian (HLNCC-FJFY-003).

Author Contributions

Lin FX, Kang DZ, Lin YX, Fang WH, Wang DL, and Chen JC participated in the study design. Zheng Y, Lin YX, He Q, Li CW, Gao ZY, and Lin FX participated in the writing of the paper. Zheng Y and Zhuo LY analyzed and explained the data. Ma K performed the database management and data cleaning. Chen RL, Huang W, Ye GZ, Wang HJ, Zhao MP, Cai LM, and Xie ZF collected the data for the study. All authors read and approved the final manuscript.

Acknowledgments

The authors thank all of the individuals who kindly participated in this study.

References

1. Kumar S, Selim MH, Caplan LR. Medical complications after stroke. Lancet Neurol. 2010 Jan; 9(1): 105–18.
2. Murthy SB, Moradiya Y, Shah J, Merkler AE, Mangat HS, Iadacola C, *et al*. Nosocomial Infections and Outcomes after Intracerebral Hemorrhage: A Population-Based Study. Neurocrit Care. 2016 Oct; 25(2): 178–84.
3. Lord AS, Lewis A, Czeisler B, Ishida K, Torres J, Kamel H, *et al.* Majority of 30-Day Readmissions After Intracerebral Hemorrhage Are Related to Infections. Stroke. 2016 Jul; 47(7): 1768–71.
4. Lord AS, Langefeld CD, Sekar P, Moomaw CJ, Badjatia N, Vashkevich A, *et al*. Infection after intracerebral hemorrhage: risk factors and association with outcomes in the ethnic/racial variations of intracerebral hemorrhage study. Stroke. 2014 Dec; 45(12): 3535–42.

5. Tinker RJ, Smith CJ, Heal C, Bettencourt-Silva JH, Metcalf AK, Potter JF, *et al.* Predictors of mortality and disability in stroke-associated pneumonia. Acta Neurol Belg. 2021 Apr; 121(2): 379–385.
6. van de Beek D, Wijdicks EF, Vermeij FH, de Haan RJ, Prins JM, Spanjaard L, *et al*. Preventive antibiotics for infections in acute stroke: a systematic review and meta-analysis. Arch Neurol. 2009 Sep; 66(9): 1076–81.
7. Westendorp WF, Vermeij JD, Zock E, Hooijenga IJ, Kruyt ND, Bosboom HJ, *et al*. The Preventive Antibiotics in Stroke Study (PASS): a pragmatic randomised open-label masked endpoint clinical trial. Lancet. 2015 Apr; 385(9977): 1519–26.
8. Kwon HM, Jeong SW, Lee SH, Yoon BW. The pneumonia score: a simple grading scale for prediction of pneumonia after acute stroke. Am J Infect Control. 2006 Mar; 34(2): 64–8.
9. Hoffmann S, Malzahn U, Harms H, Koennecke HC, Berger K, Kalic M, *et al.* Berlin Stroke Register and the Stroke Register of Northwest Germany. Development of a clinical score (A2DS2) to predict pneumonia in acute ischemic stroke. Stroke. 2012 Oct; 43(10): 2617–23.
10. Ji RJ, Shen HP, Pan YS, Wang PL, Liu GF, Wang YL, *et al.* China National Stroke Registry Investigators. Novel risk score to predict pneumonia after acute ischemic stroke. Stroke. 2013 May; 44(5): 1303–9.
11. Ji RJ, Shen HP, Pan YS, Du WL, Wang PL, Liu GF, *et al.* China National Stroke Registry investigators. Risk score to predict hospital-acquired pneumonia after spontaneous intracerebral hemorrhage. Stroke. 2014 Sep; 45(9): 2620–8.
12. Friedant AJ, Gouse BM, Boehme AK, Siegler JE, Albright KC, Monlezun DJ, *et al*. A simple prediction score for developing a hospital-acquired infection after acute ischemic stroke. J Stroke Cerebrovasc Dis. 2015 Mar; 24(3): 680–6.
13. Smith CJ, Bray BD, Hoffman A, Meisel A, Heuschmann PU, Wolfe CD, *et al.* Intercollegiate Stroke Working Party Group. Can a novel clinical risk score improve pneumonia prediction in acute stroke care? A UK multicenter cohort study. J Am Heart Assoc. 2015 Jan; 4(1): e001307.
14. Marini S, Morotti A, Lena UK, Goldstein JN, Greenberg SM, Rosand J, *et al*. Men Experience Higher Risk of Pneumonia and Death After Intracerebral Hemorrhage. Neurocrit Care. 2018 Feb; 28(1): 77–82.
15. Divani AA, Hevesi M, Pulivarthi S, Luo X, Souslian F, Suarez JI, *et al*. Predictors of nosocomial pneumonia in intracerebral hemorrhage patients: a multi-center observational study. Neurocrit Care. 2015 Apr; 22(2): 234–42.
16. Lioutas VA, Marchina S, Caplan LR, Selim M, Tarsia J, Catanese L, *et al*. Endotracheal Intubation and In-Hospital Mortality after Intracerebral Hemorrhage. Cerebrovasc Dis. 2018; 45(5–6): 270–278.
17. Chen CY, Lin WC, Yang HY. Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research. Respir Res. 2020 Feb; 21(1): 45.
18. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17; 132(20): 1920–30.

19. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018 Dec; 284(6):603–619.
20. Raju B, Jumah F, Ashraf O, Narayan V, Gupta G, Sun H, *et al*. Big data, machine learning, and artificial intelligence: a field guide for neurosurgeons. J Neurosurg. 2020 Oct: 1–11.
21. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan; 350: g7594.
22. American Thoracic Society. Infectious Diseases Society of America. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med. 2005 Feb; 171(4): 388–416.
23. Kalil AC, Metersky ML, Klompas M, Muscedere J, Sweeney DA, Palmer LB, *et al*. Management of Adults With Hospital-acquired and Ventilator-associated Pneumonia: 2016 Clinical Practice Guidelines by the Infectious Diseases Society of America and the American Thoracic Society. Clin Infect Dis. 2016 Sep; 63(5): e61-e111.
24. Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics. 2020 May; 36(10): 3093–3098.
25. Alhazmi A, Alhazmi Y, Makrami A, Masmali A, Salawi N, Masmali K, *et al*. Application of artificial intelligence and machine learning for prediction of oral cancer risk. J Oral Pathol Med. 2021 May;50(5):444–450.
26. Hanley DF, Thompson RE, Rosenblum M, Yenokyan G, Lane K, McBee N, *et al*. Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial. Lancet. 2019 Mar 9;393(10175):1021–1032.
27. Li X, Wu M, Sun C, Zhao Z, Wang F, Zheng X, *et al*. Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients. Eur J Neurol. 2020 Aug;27(8):1656–1663.
28. Urbizu A, Martin BA, Moncho D, Rovira A, Poca MA, Sahuquillo J, *et al*. Machine learning applied to neuroimaging for diagnosis of adult classic Chiari malformation: role of the basion as a key morphometric indicator. J Neurosurg. 2018 Sep; 129(3): 779–791.
29. Muhlestein WE, Akagi DS, McManus AR, Chambless LB. Machine learning ensemble models predict total charges and drivers of cost for transsphenoidal surgery for pituitary tumor. J Neurosurg. 2018 Sep 21; 131(2): 507–516.
30. Shah AA, Devana SK, Lee C, Bugarin A, Lord EL, Shamie AN, *et al*. Machine learning-driven identification of novel patient factors for prediction of major complications after posterior cervical spinal fusion. Eur Spine J. 2021 Aug 15: 10.1007/s00586-021-06961-7
31. Hotter B, Hoffmann S, Ulm L, Meisel C, Bustamante A, Montaner J, *et al*. External Validation of Five Scores to Predict Stroke-Associated Pneumonia and the Role of Selected Blood Biomarkers. Stroke. 2021 Jan; 52(1): 325–330.

32. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013 Mar 22; 14: 106.
33. Hess AS, Hess JR. Principal component analysis. Transfusion. 2018 Jul; 58(7): 1580–1582.

Figures

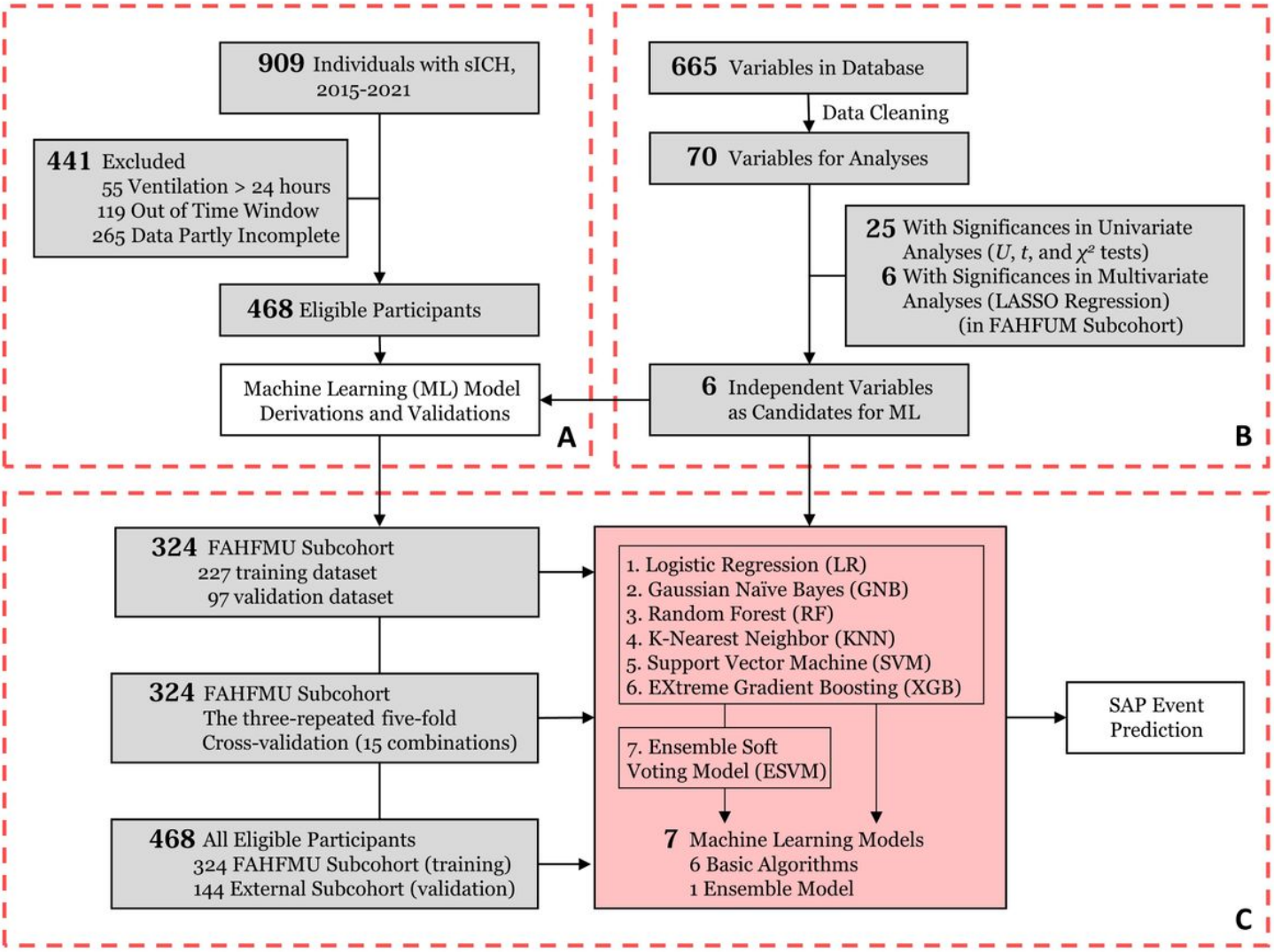

Figure 1

Flowchart of the current work. (A) Participant enrollment in the retrospective cohort of the *Risa-MIS-ICH* study; (B) Data flow from the FAHFUM subcohort; (C) The prediction model derivations and internal/cross-/external validations for stroke-associated pneumonia (SAP) events.

sICH: supratentorial intracerebral hemorrhage, LASSO: least absolute shrinkage and selection operator.

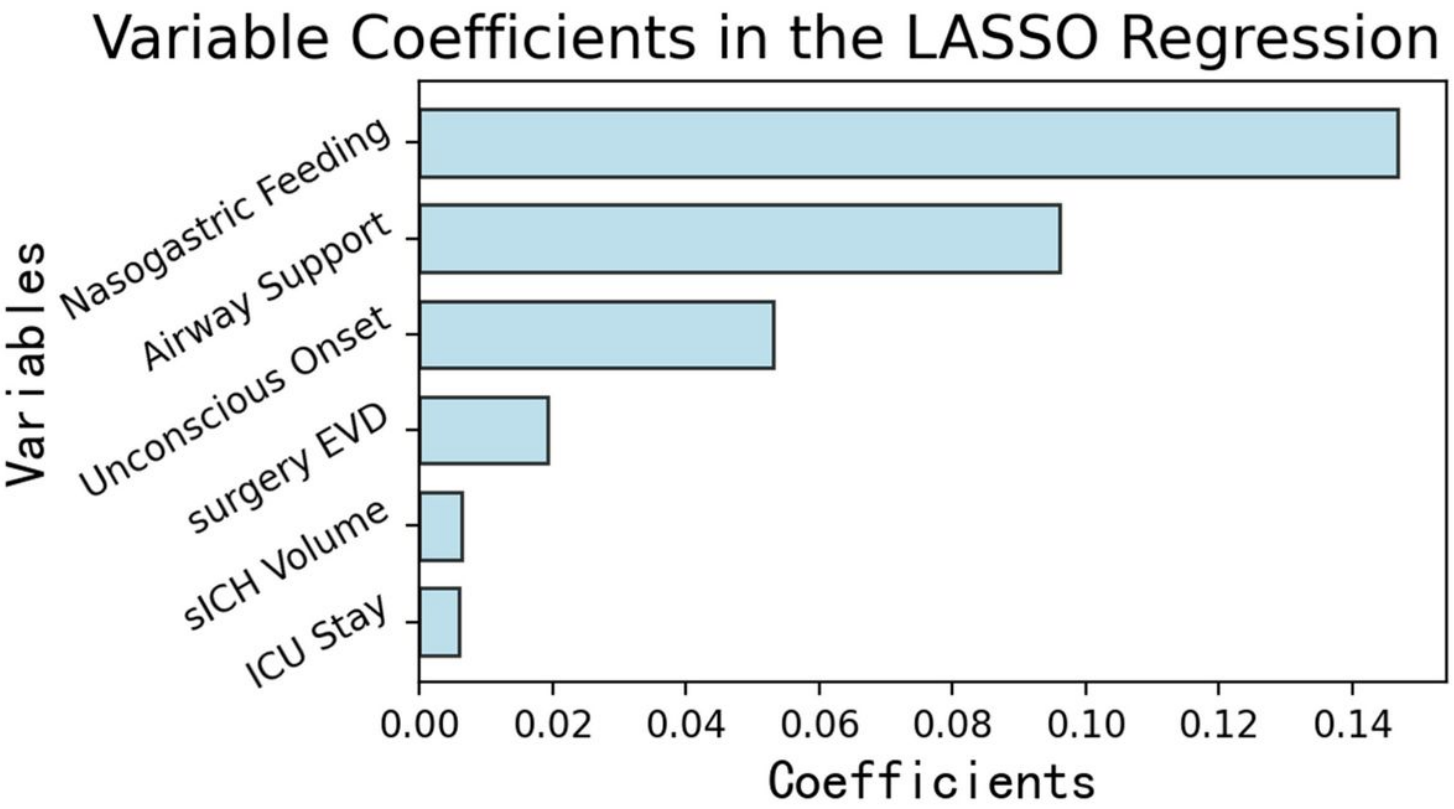

Figure 2

Importance ranking of six independent variables selected by least absolute shrinkage and selection operator (LASSO) regression: (1) nasogastric feeding, (2) airway support, (3) unconscious onset, (4) surgery for external ventricular drainage (EVD), (5) supratentorial intracerebral hemorrhage (sICH) volume, and (6) intensive care unit (ICU) stay.

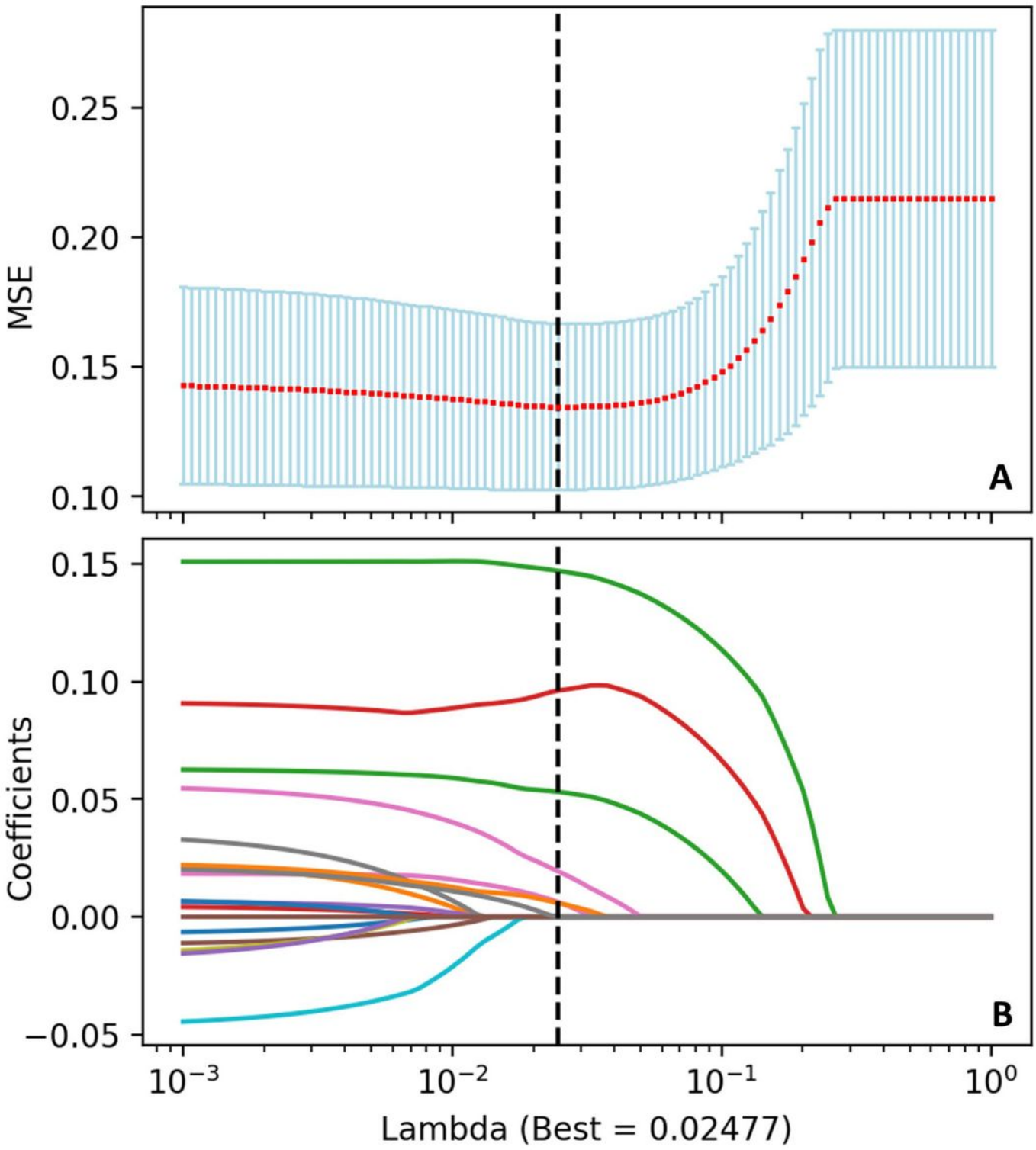

Figure 3

Multivariate analysis and variable filtrations with least absolute shrinkage and selection operator (LASSO) regression. The tuning parameter (λ) was selected for the minimized mean-square error (MSE) in the LASSO model using tenfold cross-validation. Features with nonzero coefficients were selected while the previous λ value was applied. (A) The MSE was plotted versus log λ. An optimal λ value of 0.02477 was chosen via the minimum criteria and presented as a black vertical dashed line. (B) LASSO coefficient

profiles of the features. Each colored line represents the coefficient of each feature, and six of them were selected as independent variables when λ equals 0.02477.

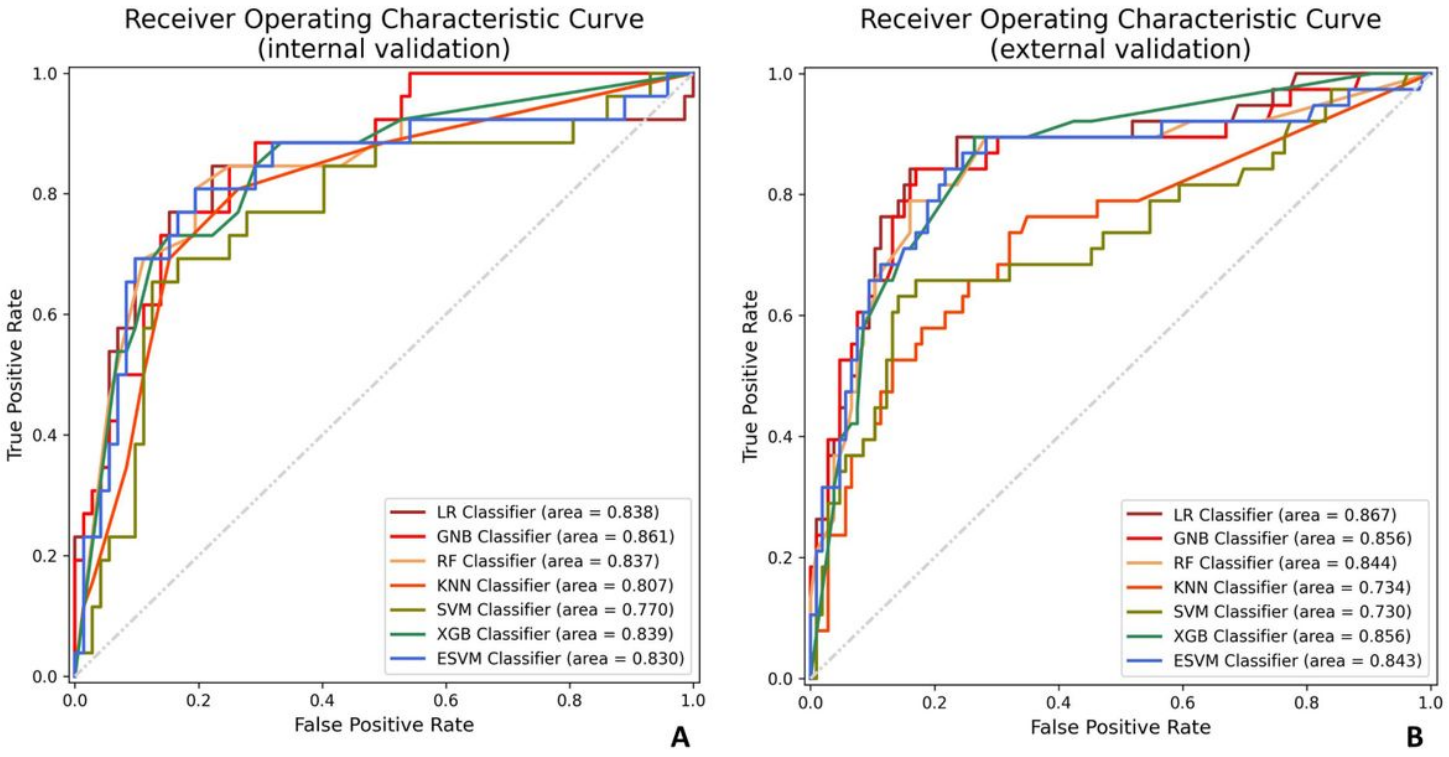

Figure 4

ROC curves for SAP on the (A) internal and (B) external validation datasets. A greater AUC value indicated a higher predictive ability of the models.

ROC: receiver operating characteristic, AUC: area under the curve, LR: logistic regression, GNB: Gaussian naïve Bayes, RF: random forest, KNN: K-nearest neighbor, SVM: support vector machine, XGB: extreme gradient boosting, ESVM: ensemble soft voting model.

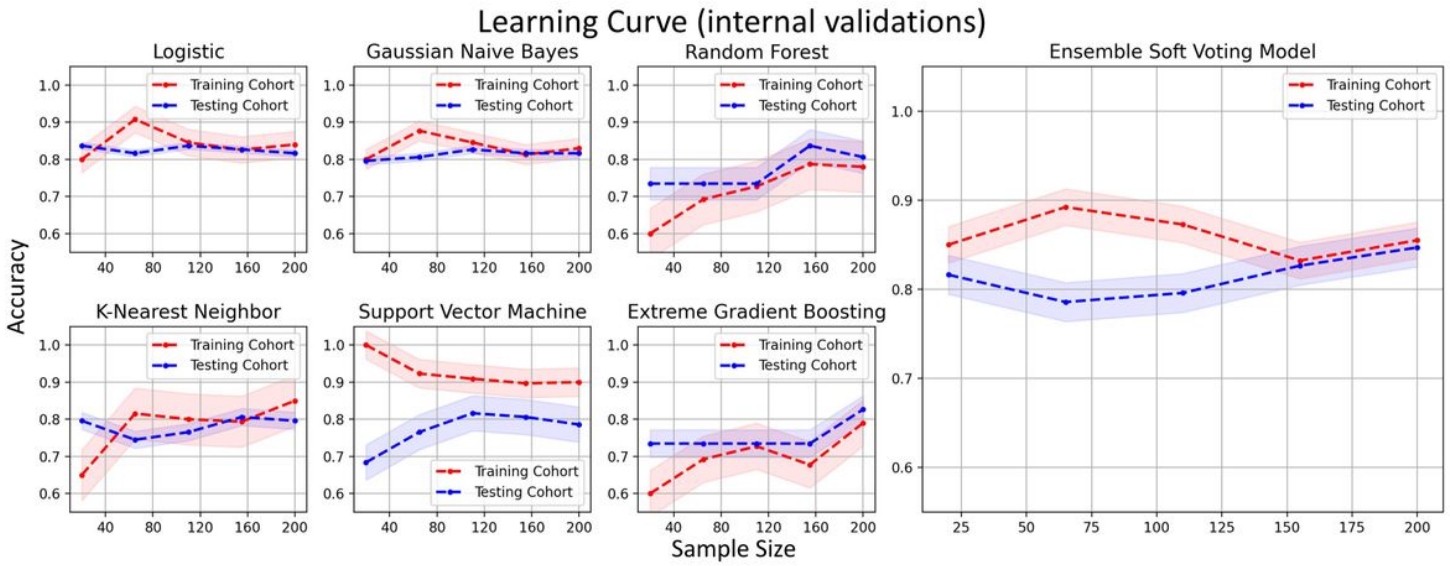

Figure 5

Learning curves of different ML prediction models in the FAHFUM subcohort. The colored area represents the 95% confidence intervals of the accuracy rates.

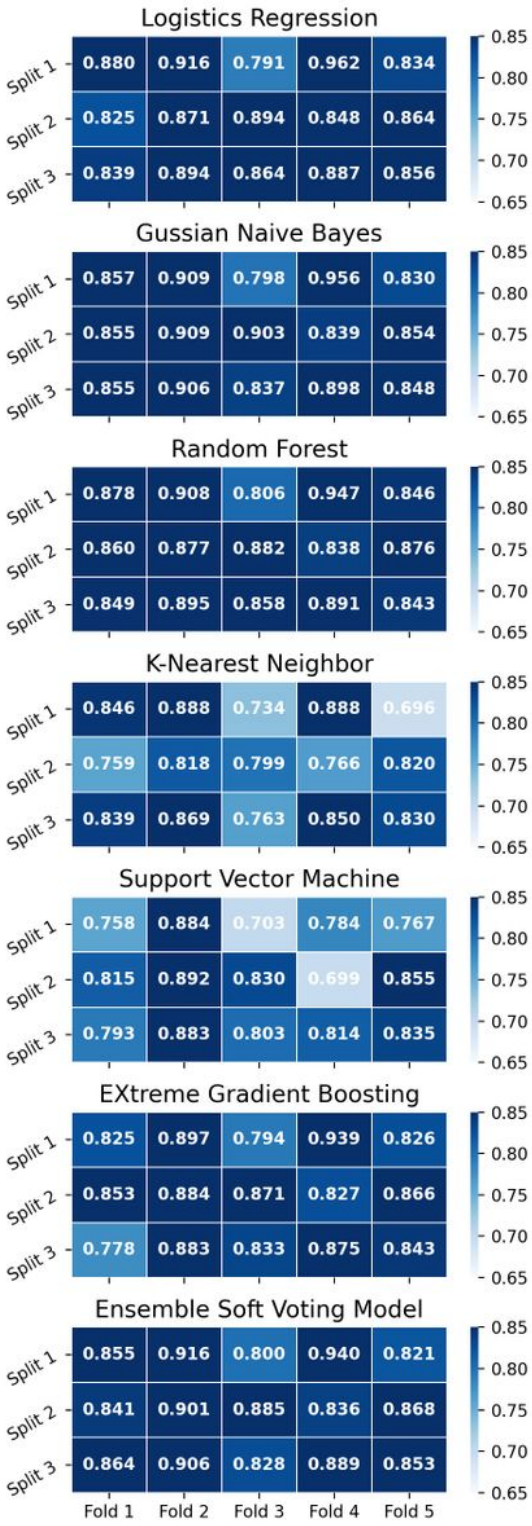

Figure 6

Heatmaps of different ML prediction models in three repeated fivefold cross-validation. Darker cells represent greater AUC values.

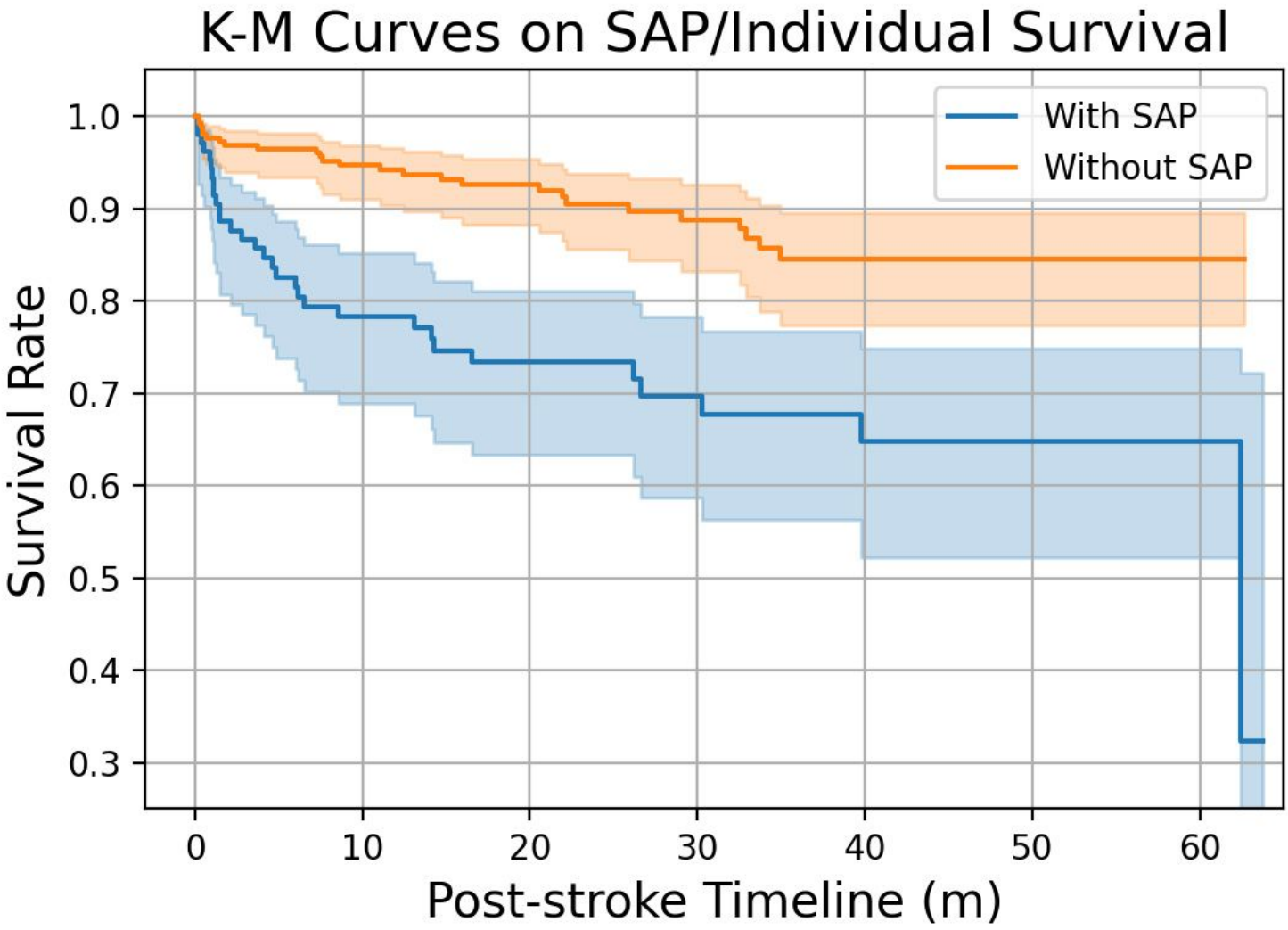

Figure 7

Kaplan–Meier curves of participants with/without SAP over 1-year follow-up. The colored area represents the 95% confidence intervals of the survival rates. The difference between both curves was examined as significant by the log-rank test (log-rank $\chi 2 = 20.34$, $P < 0.001$)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFile1.docx
- SupplementalFile2.docx