Journal of Gastroenterology and Hepatology
JGHF
doi:10.1111/jgh.16125

ORIGINAL ARTICLE - BILIARY AND PANCREATIC

Development and evaluation of machine learning models and nomogram for the prediction of severe acute pancreatitis

Zhu Luo,*[1] Jialin Shi,†[1] Yangyang Fang,‡ Shunjie Pei,‡ Yutian Lu,§ Ruxia Zhang,* Xin Ye,* Wenxing Wang,¶ Mengtian Li,* Xiangjun Li,* Mengyue Zhang,* Guangxin Xiang,‡ Zhifang Pan† and Xiaoqun Zheng*,‡,**

Departments of *Clinical Laboratory, ¶Gastroenterology and Hepatology, Second Affiliated Hospital of Wenzhou Medical University, †Key Laboratory of Intelligent Medical Imaging of Wenzhou, First Affiliated Hospital of Wenzhou Medical University, ‡School of Laboratory Medicine and Life Sciences, Wenzhou Medical University, **Key Laboratory of Laboratory Medicine, Ministry of Education of China, Wenzhou, §Department of Clinical Laboratory, Affiliated Central Hospital of Taizhou University, Taizhou, China

Key words
Machine learning, Nomogram, Prediction, Random forest model, Severe acute pancreatitis.

Accepted for publication 16 January 2023.

Correspondence
Xiaoqun Zheng, Department of Clinical Laboratory, The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou 325027, Zheijang, China.
Email: jszhengxq@163.com

Zhifang Pan, Key Laboratory of Intelligent Medical Imaging of Wenzhou, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, Zheijang, China.
Email: panzhifang@wmu.edu.cn

Financial support: The study was supported by the Wenzhou Basic Scientific Research Project (2020Y0775).
Author contributions: ZL and JLS analyzed and drafted the manuscript; YYF and SJP provided technical support; YTL, RXZ, and XY collected samples for this study; WXW, MTL, XJL, MYZ, and GXX provided the clinical data; XQZ revised it critically for important intellectual contents; XQZ and ZFP were responsible for conception and design of this study and revised the manuscript. All authors approved the final manuscript submitted for publication.

Abstract

Background and Aim: Severe acute pancreatitis (SAP) in patients progresses rapidly and can cause multiple organ failures associated with high mortality. We aimed to train a machine learning (ML) model and establish a nomogram that could identify SAP, early in the course of acute pancreatitis (AP).
Methods: In this retrospective study, 631 patients with AP were enrolled in the training cohort. For predicting SAP early, five supervised ML models were employed, such as random forest (RF), *K*-nearest neighbors (KNN), and naive Bayes (NB), which were evaluated by accuracy (ACC) and the areas under the receiver operating characteristic curve (AUC). The nomogram was established, and the predictive ability was assessed by the calibration curve and AUC. They were externally validated by an independent cohort of 109 patients with AP.
Results: In the training cohort, the AUC of RF, KNN, and NB models were 0.969, 0.954, and 0.951, respectively, while the AUC of the Bedside Index for Severity in Acute Pancreatitis (BISAP), Ranson and Glasgow scores were only 0.796, 0.847, and 0.837, respectively. In the validation cohort, the RF model also showed the highest AUC, which was 0.961. The AUC for the nomogram was 0.888 and 0.955 in the training and validation cohort, respectively.
Conclusions: Our findings suggested that the RF model exhibited the best predictive performance, and the nomogram provided a visual scoring model for clinical practice. Our models may serve as practical tools for facilitating personalized treatment options and improving clinical outcomes through pre-treatment stratification of patients with AP.

Ethical approval: The study design was approved by the Research Ethics Board of the Second Affiliated Hospital of Wenzhou Medical University and all analyses were performed following the Declaration of Helsinki (2020-K-03-01).
Guarantor of the article: Xiaoqun Zheng, PhD and Zhifang Pan, PhD.

[1]Zhu Luo and Jialin Shi contributed equally to the study.

Introduction

Acute pancreatitis (AP) is a common gastrointestinal cause for hospital admission[1] and is characterized by local and systemic inflammatory responses.[2] Globally, approximately 10%–20% of patients with AP develop severe acute pancreatitis (SAP) with organ failure or pancreatic necrosis.[3,4] For these patients, the need for intensive care is anticipated as they are at a high risk of mortality.[5] Although traditional scoring systems have been commonly used for predicting SAP, it is crucial to construct an accurate prediction model to identify patients with a high possibility of developing SAP on admission.

Many clinical scoring systems, including the Ranson score, the Glasgow score, the Bedside Index for Severity in Acute Pancreatitis (BISAP), and the Modified Computed Tomography Severity Index (MCTSI), have been employed to predict severity in patients with AP.[6] The Ranson and Glasgow scores both require 48 h ahead of application and therefore potentially miss the early window for

therapy.[7] Although BISAP is easy to execute, its sensitivity (SEN) and positive predictive value (PPV) were only 70% and 40%, respectively.[8] Computed tomography scanning is used as the reference standard for the diagnosis of AP, and MCTSI was improved; however, their values of the area under the curve (AUC) for receiver operating characteristic curve (ROC) for predicting SAP were not high.[9]

In the healthcare field, machine learning (ML) is advancing rapidly,[10] enabling computers to discover hidden medical information automatically using analytical models, which include the support vector machine, the random forest (RF), the *K*-nearest neighbors (KNN), the naive Bayes (NB), and other supervised learning, as well as unsupervised algorithms.[11] Among them, the RF model is a popular ensemble method used for classification or regression.[12] The NB model is an effective ML algorithm based on Bayes' theorem, requiring a small amount of training data, and has a simpler structure.[13] The KNN model is also widely used owing to excellent generalization and easy implementation.[14] However, to the best of our knowledge, no effective ML models for predicting SAP have been reported.

Nomogram is a graphical tool that allows for individualized risk estimation, facilitating clinical decision making.[15] Xi Cao et al.[16] established a nomogram for predicting SAP, and the results showed the external validation concordance index was 0.71 (95% confidence interval [CI]: 0.67–0.76). Guanghua Liu et al.[17] also developed a nomogram for early assessment of the SAP, and the external validation concordance index was 0.82 (95% CI [0.75, 0.89]). Thus nomogram can be used as a visual tool for predicting SAP in the clinic.

We aimed to provide and compare ML models and nomogram to predict SAP on admission using blood biomarkers and radiological parameters. The outcome may help in reducing the probability of transferring patients with AP to intensive care units with high medical costs and improving clinical outcomes.

Methods

Patients. A total of 673 patients with AP from the Second Affiliated Hospital of Wenzhou Medical University were enrolled in the training cohort between January 2015 and December 2020. A total of 109 AP patients from the Affiliated Central Hospital of Taizhou University were included in the validation cohort between January 2020 and December 2020. The diagnosis of AP was confirmed if two of the three following criteria were satisfied: upper abdominal pain, serum amylase or lipase at least three times greater than the upper limit of the normal range, and/or findings consistent with AP on imaging.[18,19] SAP was defined as patients with persistent (more than 48 h) organ failure determined using the modified Marshall score.[18,20] Patients with conditions such as chronic pancreatitis, malignant tumors, traumatic pancreatitis, anemia, pregnancy, or those under the age of 18 were excluded. The study design was approved by the Research Ethics Board of the Second Affiliated Hospital of Wenzhou Medical University, and all analyses were performed following the Declaration of Helsinki. The schematic of the patient recruitment process in the training cohort is shown in Figure 1.

The patients were divided into two groups, with one including patients with SAP and the other consisting of non-SAP individuals. Demographic characteristics of the patients, including age, gender, body mass index (BMI), and comorbidities (hypertension, diabetes, and/or hyperlipidemia) were recorded. The detailed clinical information of the patients is listed in Tables 1 and S1.

Clinical data. In this study, routine blood tests and C-reactive protein (CRP) levels from the peripheral venous blood were examined (Sysmex XI–5000 platform, Hyogo, Japan). The serum biochemical index was obtained from the examination of peripheral venous blood (Siemens ADVIA 2400 system, Nurnberg, Germany). Laboratory examinations were performed on admission and/or after 48 h, while the computed tomography scanning was performed only on admission. The BISAP and MCTSI scores were calculated on admission, and the Ranson and Glasgow scores were estimated from the data of the first 48 h post-admission.

For the convenience of clinical application, while developing prediction models, continuous factors were turned into categorical ones based on the cut-off values. And the cut-off values of all these variables were obtained by ROC curves, with the values reaching the maximum AUC (Youden's index = sensitivity + specificity − 1). As described by Hu et al.,[21] the ROC curve and Youden's index were also calculated to identify the best predictor of metabolic abnormalities in Chinese adults. The cut-off values were as follows: CRP ≥ 95 mg/dL, white blood cell count (WBC) ≥ 16.5×10^9 L, neutrophil to lymphocyte ratio (NLR) ≥ 10, alanine aminotransferase (ALT) ≥ 61.5 U/L, aspartate aminotransferase (AST) ≥ 43.5 U/L, albumin (ALB) < 30 g/L, glucose (GLU) ≥ 9 mmol/L, blood urea nitrogen (BUN) > 20 mg/dL, calcium (CA) ≥ 2 mmol/L, high-density lipoprotein (HDL) ≥ 0.7 mmol/L, pancreatic inflammation (PI) (yes vs. no), pancreatic necrosis (PN) (yes vs. no), and extrapancreatic complications (yes vs. no). In addition, the cut-off for creatinine (CR) ≥ 134 μmol/L (1.5 mg/dl) was based on previously published data.[22,23]

ML models and nomogram. We chose five ML models, including the RF, NB, KNN, neural network (NN), and classification tree (CT) as ML classifiers for predicting SAP. We do some preprocessing to the collected data, such as using smote algorithm to process the unbalanced samples, expanding the new samples, and generating new data. In both the training and validation cohorts, we randomly divided the data into the "train set" and "test set" in a 7:3 ratio. The train set was used for training the model by 10-fold cross-validation, following which its predictive performance was evaluated in the test set. In this model, 1000 random samples are generated by replacement to determine how many features are included in the model at least and compared with the existing scoring system. In addition, the relatively important factors were filtered out by a bootstrap analysis based on the RF model in the training cohort.

To overcome the black box problem of ML models output and improve its interpretability, the nomogram was set up based on the coefficients of LRM and used to explain the individualized prediction. Meanwhile, the nomogram was evaluated by the calibration curve and AUC.

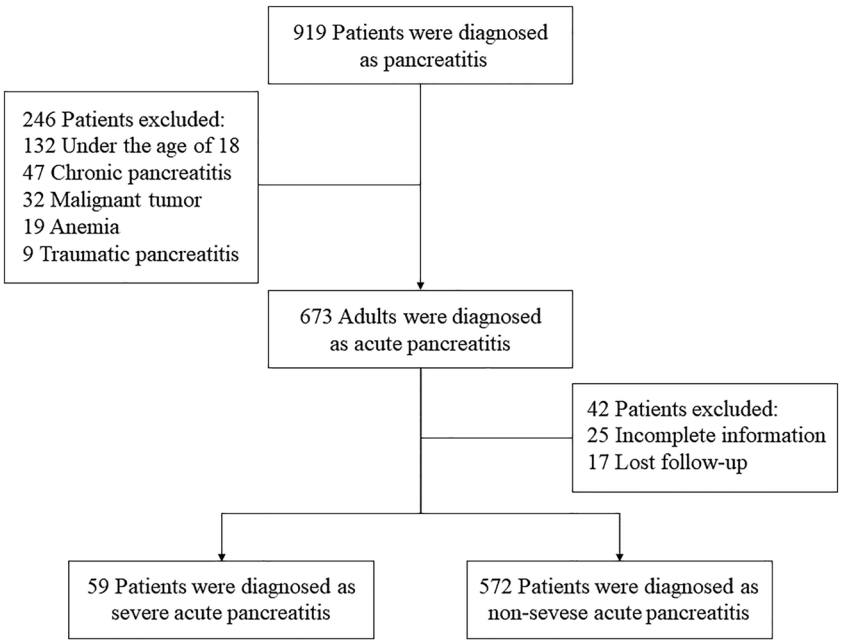

Figure 1 Patient recruitment process in the training cohort. In total, 631 out of 919 patients were included in the training cohort according to the selection criteria. The included patients underwent laboratory tests and computed tomography scanning. Their complete clinical information was available and included in the study.

Statistical analysis. Statistical analysis was performed using the SPSS software (version 25.0) (Chicago, IL) and R statistical software (version 4.0.5). The categorical data were compared using the χ^2 test or the Fisher exact test, while the continuous variables were compared by *t*-test or the Mann–Whitney *U* test. The results were presented as percentages (%), mean ± standard deviation, or median (25th–75th percentile). Univariate and multivariate regression analyses were performed to evaluate the influence of clinical characteristics. The AUC, accuracy (ACC), SEN, specificity (SPE), PPV, and negative predictive value (NPV) measures were employed to compare the models with the existing traditional scoring systems for predicting severity, and ROC curves were plotted on GraphPad Prism (version 8.0.1) (San Diego, CA, USA). The nomogram and calibration curve were built on R statistical software (version 4.0.5). Results were expressed as odds ratio (OR) with the corresponding 95% confidence interval (CI). $P < 0.05$ was considered statistically significant.

Results

Clinical characteristics and risk factors for patients with AP in the training and validation cohorts. Distributions of clinical characteristics and risk factors between the two study cohorts were listed in Table 1 and Table S1. In the training and validation cohorts, 631 and 109 patients were enrolled, respectively. The percentage of SAP was 9.3% and 9.1% in the training and validation cohorts, respectively. There were no statistically significant differences in age, gender, comorbidities, and BMI between the SAP and non-SAP groups. However, the length of hospital stay in the SAP group was significantly longer than that in the non-SAP group ($P < 0.001$).

Univariate analysis in the training cohort. A total of 20 variables considered associated with SAP were verified by univariate analysis in the training cohort (Table S2). CRP, WBC, NLR, ALT, AST, ALB, GLU, BUN, CR, CA, HDL, PI, PN, and extrapancreatic complications were identified as candidate variables of SAP, while no statistical differences were observed for other clinical features.

Performances of different ML models in the training cohort. We incorporated the 14 candidate variables (categorical) into five ML models including RF, KNN, CT, NB, and NN, which were used to predict SAP early in the course of AP in the training cohort. The result showed that the ACC and AUC for RF, KNN, CT, NB, and NN for predicting SAP were 90.1% and 0.969 (95% CI [0.953, 0.984]), 88.6% and 0.954 (95% CI [0.934, 0.975]), 84.9% and 0.892 (95% CI [0.856, 0.927]), 90.1% and 0.951 (95% CI [0.928, 0.974]), and 88.9% and 0.932 (95% CI [0.903, 0.961]), respectively (Table 2). In addition, the AUC for the traditional scoring systems, including MCTSI, BISAP, Ranson, Glasgow scores, and SIRS for predicting SAP were 0.801 (95% CI [0.736, 0.865]), 0.796 (95% CI [0.744, 0.849]), 0.847 (95% CI [0.798, 0.900]), 0.837 (95% CI [0.792, 0.882]), and 0.743 (95% CI [0.686, 0.800]), respectively. Through comprehensive comparison, the RF, KNN, and NB models showed superior performances for predicting SAP as evidenced by the high AUC, ACC, SEN, SPE, PPV, and NPV values (Fig. 2a).

Performances of the RF, KNN, and NB models in the validation cohort. Three ML models including RF, KNN, and NB were then used to predict SAP in the validation cohort. The ACC and AUC for RF, KNN, and NB models for the prediction of SAP were 86.0% and 0.961 (95% CI [0.920, 1.000]), 80.7% and 0.947 (95% CI [0.893, 1.000]), 75.4% and

Table 1 Clinical characteristics and risk factors for patients with AP in the training and validation cohorts

Variables	Training cohort (*n* = 631)	Validation cohort (*n* = 109)
Age (y, IQR)	44 (35–56)	43 (33–62)
Male (*n*, %)	428 (67.8%)	63 (57.8%)
Comorbidities (*n*, %)	252 (39.9%)	42 (38.5%)
Body mass index (kg/m^2, IQR)	23.8 (22.0–26.5)	24.9 (21.5–27.7)
Hospital stay (d, IQR)	9 (6–15)	8 (6–10)
C-reactive protein (mg/dL, IQR)	42 (10–118)	70 (14–123)
White blood cell count (×10^{9}L, IQR)	11.4 (8.7–14.4)	10.8 (7.0–13.6)
Neutrophil ratio (%, IQR)	0.83 (0.75–0.88)	0.81 (0.74–0.86)
Lymphocyte ratio (%, IQR)	0.11 (0.07–0.17)	0.12 (0.08–0.18)
Hemoglobin (g/L, IQR)	144 (130–157)	136 (122–154)
RBC (×10^{12}L, IQR)	4.7 (4.3–5.1)	4.5 (4.1–5.0)
Hematocrit (IQR)	0.43 (0.39–0.46)	0.41 (0.37–0.45)
Red cell distribution width (%, IQR)	13.0 (12.5–13.6)	12.9 (12.6–13.3)
PLT (×10^{9}L, IQR)	206 (166–252)	194 (153–242)
Platelet distribution width (%, IQR)	15.9 (12.9–16.7)	16.3 (16.0–16.7)
Alanine aminotransferase (U/L, IQR)	31 (17–69)	23 (15–41)
Aspartate aminotransferase (U/L, IQR)	26 (17–48)	22 (17–33)
Albumin (g/L, IQR)	38.8 (35.5–41.7)	39.3 (36.0–41.8)
Glucose (mmol/L, IQR)	7.3 (5.8–9.7)	7.6 (6.3–10.4)
Blood urea nitrogen (mmol/L, IQR)	5.0 (3.8–6.6)	4.1 (3.3–5.1)
Creatinine (umol/L, IQR)	61 (50–72)	64 (57–72)
Lactate dehydrogenase (IU/L, IQR)	222 (177–382)	214 (186–274)
Calcium (mmol/L, IQR)	2.1 (2.0–2.2)	2.1 (2.0–2.2)
Triglycerides (mmol/L, IQR)	1.7 (0.9–5.5)	1.5 (0.9–7.0)
Total cholesterol (mmol/L, IQR)	4.9 (3.9–6.1)	5.1 (4.1–6.5)
High-density lipoprotein (mmol/L, IQR)	0.9 (0.7–1.2)	0.9 (0.7–1.1)
Low-density lipoprotein (mmol/L, IQR)	2.1 (1.3–2.9)	2.0 (1.2–3.0)
Amylase (IU/L, IQR)	288 (116–753)	220 (80–476)
Lipase (IU/L, IQR)	240 (105–630)	175 (73–417)
Computed tomography scanning		
Pancreatic inflammation (*n*, %)	544 (86.2%)	107 (98.2%)
Pancreatic necrosis (*n*, %)	13 (2.1%)	2 (1.8%)
Extrapancreatic complications (*n*, %)	199 (31.5%)	9 (8.3%)
MCTSI (IQR)	2 (2–4)	1 (1–1)
BISAP (IQR)	1 (0–1)	1 (0–1)
Ranson (IQR)	1 (0–1)	0 (0–1)
Glasgow (IQR)	0 (0–1)	1 (0–1)
SIRS (IQR)	1 (0–1)	1 (0–2)

Note: Data as mean ± standard deviation, or numbers and percentages, or median (25th–75th percentile), as appropriate.
Abbreviations: BISAP, Bedside Index for Severity in Acute Pancreatitis; IQR, interquartile range; MCTSI, Modified Computed Tomography Severity Index; *N*, number; SIRS, Systemic Inflammatory Response Syndrome.

Table 2 Comparison of performances of different models in the training cohort

Models	AUC	ACC%	SEN%	SPE%	PPV%	NPV%	*P* value
RF	0.969 [0.953–0.984]	90.1	88.6	91.5	91.2	89.0	<0.001
KNN	0.954 [0.934–0.975]	88.6	89.7	87.6	87.7	89.6	<0.001
CT	0.892 [0.856–0.927]	84.9	83.4	86.4	85.9	84.1	<0.001
NB	0.951 [0.928–0.974]	90.1	90.3	89.8	89.8	90.3	<0.001
NN	0.932 [0.903–0.961]	88.9	88.0	89.8	89.5	88.3	<0.001
MCTSI	0.801 [0.736–0.865]	50.7	88.1	69.1	22.7	98.3	<0.001
BISAP	0.796 [0.744–0.849]	55.3	62.7	86.0	64.0	92.9	<0.001
Ranson	0.847 [0.798–0.900]	51.0	76.3	81.8	47.1	95.2	<0.001
Glasgow	0.837 [0.792–0.882]	64.7	72.9	87.8	58.5	94.1	<0.001
SIRS	0.743 [0.686–0.800]	46.8	61.0	80.4	24.3	95.2	<0.001

Note: 95% confidence intervals were included in brackets.
Abbreviations: ACC, accuracy; AUC, area under the curve; BISAP, Bedside Index for Severity in Acute Pancreatitis; CT, classification tree; KNN, *K*-nearest neighbors; MCTSI, Modified Computed Tomography Severity Index; NB, native Bayes; NN, neural network; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SEN, sensitivity; SIRS, Systemic Inflammatory Response Syndrome; SPE, specificity.

0.954 (95% CI [0.906, 1.000]), respectively (Table 3, Fig. 2b). Although the AUC, SEN, and NPV of the NB model were similar to those of the RF model, the SPE and PPV of the former were only 48.2% and 68.2%, thereby increasing the chances of clinical misdiagnoses. Therefore, the RF model in the validation cohort, along with the above training cohort, exhibited the best performance for predicting SAP.

Ranking the important candidate variables based on the RF model. We ranked the importance of these 14 candidate variables based on the RF prediction model for SAP in the training and validation cohorts (Fig. 2C). The relatively important variables were identified, including extrapancreatic complications, AST, CA, BUN, ALB, WBC, and NLR. They were further determined as predictive factors for SAP with standard box plots shown in Figure S3.

Nomogram based on the predictive factors. Among these aforementioned 14 candidate variables (categorical), 7 variables were found to be relevant for SAP in the multivariate analysis (Table S4), namely, extrapancreatic complications (OR 5.495, 95% CI [2.320, 13.013]), AST (OR 3.673, 95% CI [1.785, 7.558]), CR (OR 3.290, 95% CI [1.644, 6.585]), CRP (OR 3.045, 95% CI [1.407, 6.587]), ALB (OR 0.412, 95% CI [0.196, 0.866]), WBC (OR 3.316, 95% CI [1.519, 7.237]), and GLU (OR 2.898, 95% CI [1.383, 6.070]). Therefore, these seven variables were retained in the LRM. The LRM was: log (odds of SAP) = 1.704 (EC) + 1.301 (AST) + 1.191 (CR) + 1.113 (CRP)-0.886 (ALB) + 1.199 (WBC) + 1.064 (GLU).

The nomogram was constructed by using the coefficients of LRM (Fig. 3a). The AUC of the nomogram was 0.888 (95% CI [0.881, 0.918]) in the training cohort and 0.955 (95% CI [0.893, 1.000]) in the validation cohort (Fig. 3b). The calibration curves

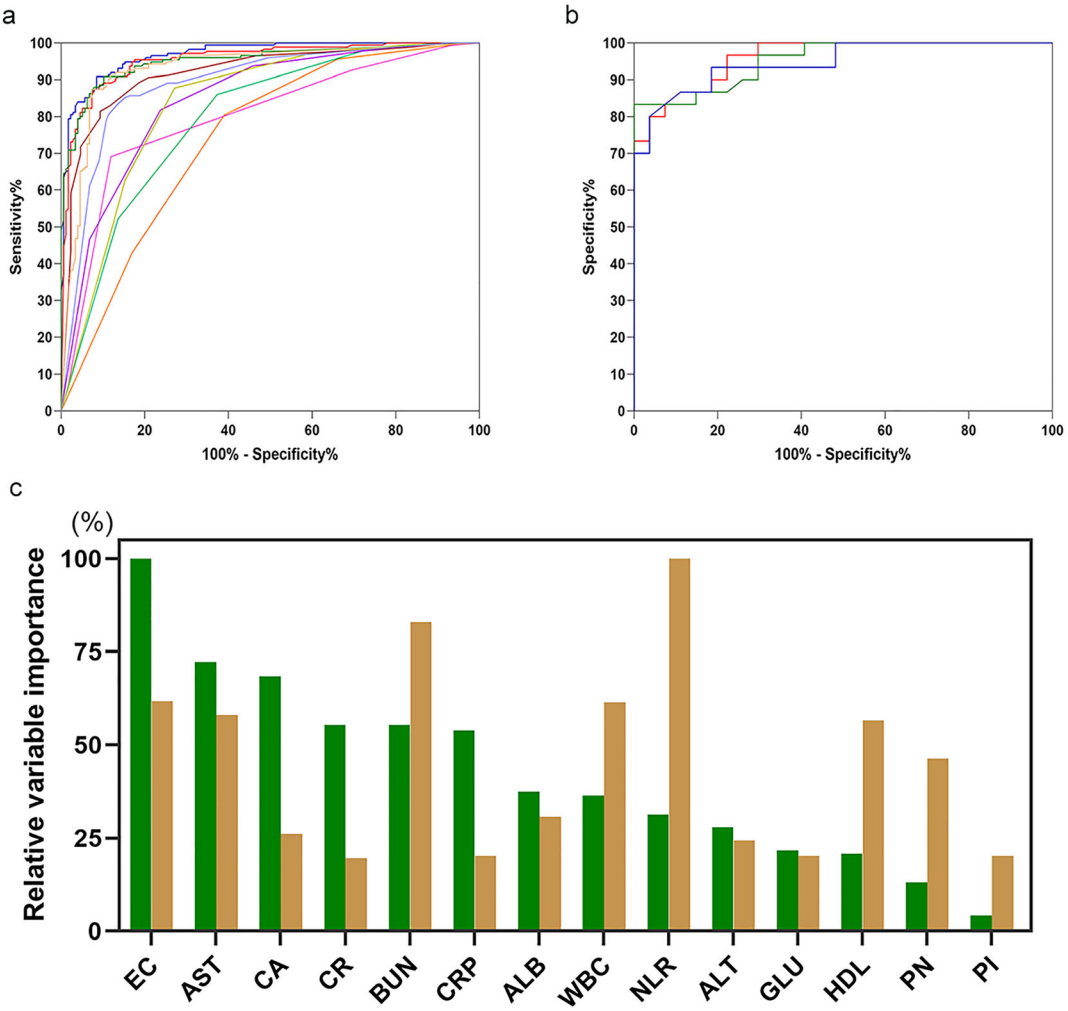

Figure 2 (a) Comparison of receiver operating characteristic (ROC) curves between different models in the training cohort. Numbers in parentheses indicated the areas under the curve (AUC) for ROC. (b) Comparison of ROC curves between random forest, *K*-nearest neighbors, and naive Bayes in the validation cohort. (c) Ranking the important variables based on the RF prediction models. Variable importance was presented as a percentage of the highest value. BISAP, Bedside Index for Severity in Acute Pancreatitis; EC, extrapancreatic complications; MCTSI, Modified Computed Tomography Severity Index; PI, pancreatic inflammation; PN, pancreatic necrosis; SIRS, Systemic Inflammatory Response Syndrome. (a) —, random forest, (0.969); —, *K*-nearest neighbors (0.954); —, naive Bayes (0.951); —, neural network (0.932); —, classification tree (0.892); —, Ranson (0.847); —, Glasgow (0.837); —, MCTSI (0.801); —, BISAP (0.796); —, SIRS (0.743). (b) —, random forest, (0.961); —, *K*-nearest neighbors (0.947); —, Naive Bayes (0.954). (c) ▬, Training cohort; ▬, validation cohort.

were shown in Figure 3c,d, which indicated a close consistency between the predicted and observed probability.

Discussion

As SAP in patients is related to multiple organ failures with a high mortality rate, we employed five ML models (RF, KNN, CT, NB, and NN) for the early determination of SAP. By comparing their prediction performances both in the training and validation cohorts, the RF model demonstrated the best performance and could be useful for guiding treatment and improving clinical outcomes.

ML algorithms based on AI technology for prediction and diagnosis have been generally accepted in the medical field.[24] The RF model performs bootstrap aggregation of multiple regression trees to reduce over-fitting and summarizes the prediction results from individual trees, thereby yielding more accurate predictions.[25,26] Previous studies also showed that RF could efficiently account for nonlinear effects and correlated parameters interactions.[27] Lan et al.[28] reported that RF was used to predict the timing of surgical intervention for infected necrotizing pancreatitis patients although they did not stratified the risk factors. Holodinsky JK et al.[29] showed the RF model was a useful ML method for predicting 90-day hometime in individuals with stroke. Yang et al.[30] used ML models to predict the risk of cardiovascular diseases and found RF superior with an AUC of 0.787. In our study, we incorporated 14 candidate variables into RF, KNN, CT, NB, and NN, and found RF to fulfill the prediction of SAP early in the course of AP.

Table 3 Comparison of performances of RF, KNN, and NB in the validation cohort

Models	AUC	ACC%	SEN%	SPE%	PPV%	NPV%	*P* value
RF	0.961 [0.920–1.000]	86.0	90.0	81.5	84.4	88.0	<0.001
KNN	0.947 [0.893–1.000]	80.7	93.3	66.7	75.7	90.0	<0.001
NB	0.954 [0.906–1.000]	75.4	100.0	48.2	68.2	100.0	<0.001

Note: 95% confidence intervals were included in brackets.
Abbreviations: ACC, accuracy; AUC, area under the curve; BISAP, Bedside Index for Severity in Acute Pancreatitis; CT, classification tree; KNN, *K*-nearest neighbors; MCTSI, Modified Computed Tomography Severity Index; NB, native Bayes; NN, neural network; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SEN, sensitivity; SIRS, Systemic Inflammatory Response Syndrome; SPE, specificity.

Valverde-López et al.[8] showed the Ranson score with an AUC of 0.85 and PPV of 22.1% for predicting SAP. Pando et al.[31] reported the BISAP score with an AUC of 0.873 and SPE of 61.6% and the Acute Physiology and Chronic Health Evaluation-II score with an AUC of 0.761 and SPE of 63.3% in predicting SAP. In our study, the AUC, ACC, SEN, SPE, PPV, and NPV for the RF model were 0.969 (0.953–0.984), 90.1%, 88.6%, 91.5%, 91.2%, and 89.0% in the training cohort (Table 2), and 0.961 (0.920–1.000), 86.0%, 90.0%, 81.5%, 84.4%, and 88.0% in the validation cohort (Table 3), respectively. Thus, compared with other scoring systems, our RF model was much more accurate in predicting SAP early in the course of AP (Fig. 2a,b). Then we ranked the variables and obtained seven significant factors, namely, extrapancreatic complications, AST, CA, BUN, ALB, WBC, and NLR (Fig. 2c).

To facilitate clinical application, the nomogram was constructed by using the coefficients of LRM, which provided a visual scoring model for the clinic (Fig. 3a). Meanwhile, we also obtained seven important variables in the nomogram, namely, extrapancreatic complications, AST, CR, CRP, ALB, WBC, and GLU. Four of the seven variables were the same as the seven important variables of the RF model, and they were extrapancreatic complications, AST, ALB, and WBC, which could be tested on admission with high efficiency, simple operability, and low cost. However, in previous studies, extrapancreatic complications were

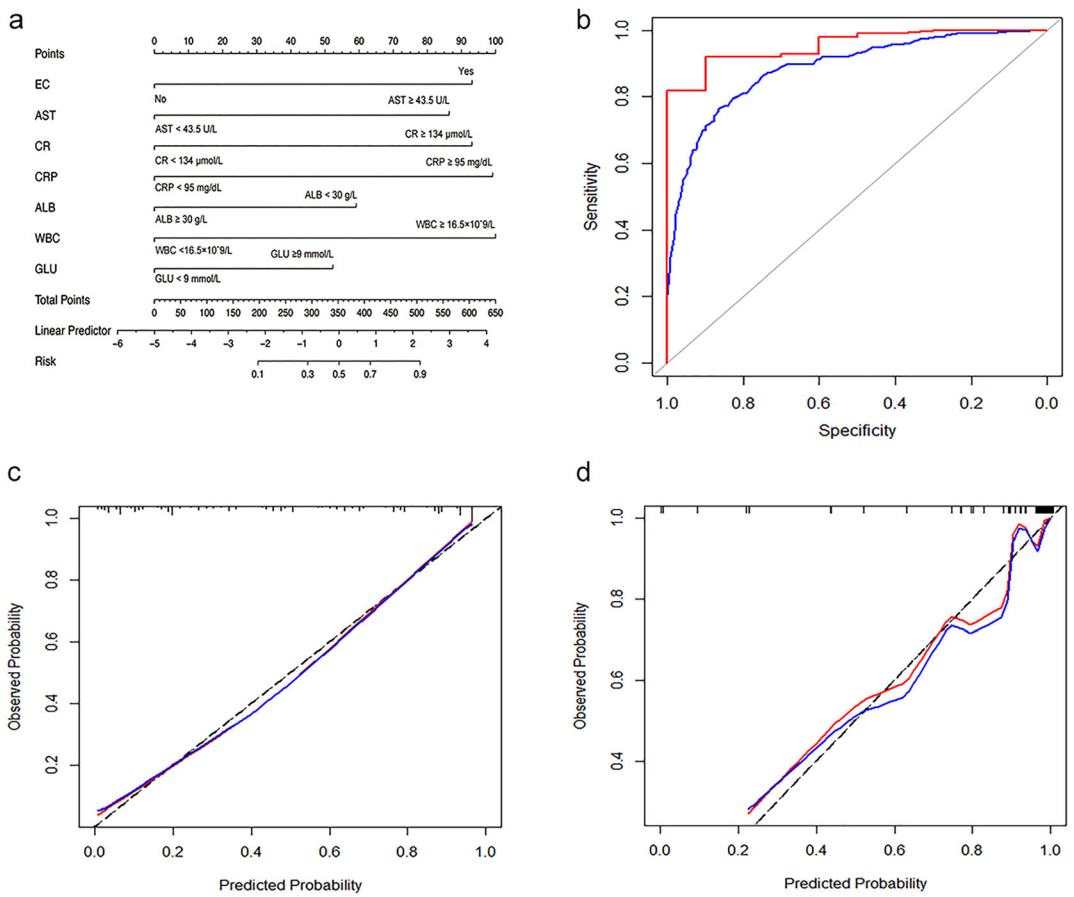

Figure 3 (a) Nomogram predicting the individual probability of severe acute pancreatitis (SAP). The points of seven variables were added to obtain the sum of points, leading to evaluate the individual probability of progression to SAP in patients with acute pancreatitis (AP). (b) The receiver operating characteristic (ROC) curves of the LRM in the training and validation cohorts. The calibration curves in the training cohorts (c) and validation cohorts (d). (b) –, Training cohort, AUC = 0.888; –, validation cohort, AUC = 0.955.

rarely included along with laboratory indicators in the clinical predictive models for SAP. Interestingly, by incorporating extrapancreatic complications into our RF model and nomogram, the disease progression of AP patients was comprehensively evaluated.

The clinical significance of different prediction models varies, and they possess their own advantages. The nomogram provided a visual and convenient prediction model for clinical practice. It can be used to predict the probability of each AP patient progressing to SAP, optimizing their early individualized treatment plan. As one of machine learning models, the RF model can deal with high-dimensional data and judge the interaction between different features. Therefore, in future research on the SAP prediction model, more dimensional variables can be easily incorporated into the RF model, keeping the SAP prediction model up-to-date and improving its prediction performance. Similar to the amount of data in our study, Chen et al.[32] established and analyzed an RF-based disease risk prediction model for the systemic lupus erythematosus, with the training and validation datasets 405 and 173, respectively. Taken together, the RF model can be used as a new auxiliary tool for disease risk prediction in clinical application and contribute to the early identification of diseases.

This study still has limitations, as clinical data and radiological parameters were not comprehensive enough. A further investigation is warranted to expand the clinical sample size, as well as the prospective and multicenter scope of the studies to consolidate RF performances. Prognosis prediction of SAP will also be of interest, and RF should be considered for a wide range of clinical applications.

In conclusions, we established and externally validated ML models and nomogram for early prediction of SAP. The RF model exhibited the best performance among ML models, with higher prediction accuracy, and supplied a feature importance ranking. Meanwhile, the nomogram explained the individualized prediction, exerted strong calibration ability, and provided a visual and convenient scoring model for the clinic. The results showed that both models showed strengths in predicting SAP, and we believe this combination might be more clinically useful than the RF model or nomogram alone. During the rapid clinical progression, early prediction of SAP in patients with AP is of great significance for timely treatment, personalized management, and improved outcomes.

Data availability statement. The data generated or analyzed during this study are available from the corresponding author on reasonable request.

References

1 Barreto SG, Habtezion A, Gukovskaya A *et al*. Critical thresholds: key to unlocking the door to the prevention and specific treatments for acute pancreatitis. *Gut* 2021; **70**: 194–203.
2 Boxhoorn L, Voermans RP, Bouwense SA *et al*. Acute pancreatitis. *Lancet* 2020; **396**: 726–34.
3 Garg PK, Singh VP. Organ failure due to systemic injury in acute pancreatitis. *Gastroenterology* 2019; **156**: 2008–23.
4 Schepers NJ, Bakker OJ, Besselink MG *et al*. Impact of characteristics of organ failure and infected necrosis on mortality in necrotising pancreatitis. *Gut* 2019; **68**: 1044–51.
5 van Dijk SM, Hallensleben NDL, van Santvoort HC *et al*. Acute pancreatitis: recent advances through randomised trials. *Gut* 2017; **66**: 2024–32.
6 Mounzer R, Langmead CJ, Wu BU *et al*. Comparison of existing clinical scoring systems to predict persistent organ failure in patients with acute pancreatitis. *Gastroenterology* 2012; **142**: 1476–82.
7 Silva-Vaz P, Abrantes AM, Castelo-Branco M, Gouveia A, Botelho MF, Tralhão JG. Multifactorial scores and biomarkers of prognosis of acute pancreatitis: applications to research and practice. *Int. J. Mol. Sci.* 2020; **21**: 338.
8 Valverde-López F, Matas-Cobos AM, Alegría-Motte C, Jiménez-Rosales R, Úbeda-Muñoz M, Redondo-Cerezo E. BISAP, RANSON, lactate and others biomarkers in prediction of severe acute pancreatitis in a European cohort. *J. Gastroenterol. Hepatol.* 2017; **32**: 1649–56.
9 Mikó A, Vigh É, Mátrai P *et al*. Computed tomography severity index vs. other indices in the prediction of severity and mortality in acute pancreatitis: a predictive accuracy meta-analysis. *Front. Physiol.* 2019; **10**: 1002.
10 Peiffer-Smadja N, Rawson TM, Ahmad R *et al*. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin. Microbiol. Infect.* 2020; **26**: 584–95.
11 Zhou Y, Ge YT, Shi XL *et al*. Machine learning predictive models for acute pancreatitis: a systematic review. *Int. J. Med. Inform.* 2022; **157**: 104641.
12 Wang Y, Xia ST, Tang Q, Wu J, Zhu X. A novel consistent random forest framework: bernoulli random forests. *IEEE Trans Neural Netw Learn Syst.* 2018; **29**: 3510–23.
13 Zhang N, Wu L, Yang J, Guan Y. Naive Bayes bearing fault diagnosis based on enhanced independence of data. *Sensors (Basel).* 2018; **18**: 463.
14 Garcia-Pedrajas N, Romero Del Castillo JA, Cerruela-Garcia G. A proposal for local *k* values for *k*-nearest neighbor rule. *IEEE Trans Neural Netw Learn Syst.* 2017; **28**: 470–5.
15 Park SY. Nomogram: an analogue tool to deliver digital knowledge. *J. Thorac. Cardiovasc. Surg.* 2018; **155**: 6.
16 Cao X, Wang HM, Lu R *et al*. Establishment and verification of a nomogram for predicting severe acute pancreatitis. *Eur. Rev. Med. Pharmacol. Sci.* 2021; **25**: 1455–61.
17 Liu GH, Chen J, Li LQ, Huan XS, Lei P. Development and validation of a nomogram for early assessment the severity of acute pancreatitis. *Scand. J. Gastroenterol.* 2022; **57**: 990–5.
18 Banks PA, Bollen TL, Dervenis C *et al*. Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus. *Gut* 2013; **62**: 102–11.
19 Forsmark CE, Vege SS, Wilcox CM. Acute pancreatitis. *N. Engl. J. Med.* 2016; **375**: 1972–81.
20 Abu Omar Y, Attar BM, Agrawal R *et al*. Revised Marshall score: a new approach to stratifying the severity of acute pancreatitis. *Dig. Dis. Sci.* 2019; **64**: 3610–5.
21 Hu J, Jiang Y, Shen H, Ding L, Xu X, Wu W *et al*. What is the best anthropometry index to evaluate the risk of metabolic abnormalities in Chinese adults? *Diabetes Metab. Res. Rev.* 2022; **38**(8):17.
22 Lin KY, Zheng WP, Bei WJ *et al*. A novel risk score model for prediction of contrast-induced nephropathy after emergent percutaneous coronary intervention. *Int. J. Cardiol.* 2017; **230**: 402–12.
23 Vaz NF, da Cunha VNR, Cunha-Silva M, Sevá-Pereira T, de Souza Almeida JR, Mazo DF. Evolution of diagnostic criteria for acute kidney injury in patients with decompensated cirrhosis: a prospective study in a tertiary university hospital. *Clin. Res. Hepatol. Gastroenterol.* 2020; **44**: 551–63.
24 Kawakami E, Tabata J, Yanaihara N *et al*. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in

epithelial ovarian cancer based on blood biomarkers. *Clin. Cancer Res.* 2019; **25**: 3006–15.
25 Breiman L. Random forests. *Machine Learning.* 2001; **45**: 5–32.
26 Van der Heide EMM, Veerkamp RF, Van Pelt ML, Kamphuis C, Athanasiadis I, Ducro BJ *et al*. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. *J. Dairy Sci.* 2019; **102**: 9409–21.
27 Ingrisch M, Schöppe F, Paprottka K *et al*. Prediction of (90)Y radioembolization outcome from pretherapeutic factors with random survival forests. *J. Nucl. Med.* 2018; **59**: 769–73.
28 Lan L, Guo Q, Zhang Z *et al*. Classification of infected necrotizing pancreatitis for surgery within or beyond 4 weeks using machine learning. *Front. Bioeng. Biotechnol.* 2020; **8**: 541.
29 Holodinsky JK, Yu AYX, Kapral MK, Austin PC. Using random forests to model 90-day hometime in people with stroke. *BMC Med. Res. Methodol.* 2021; **21**: 102.
30 Yang L, Wu H, Jin X *et al*. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci. Rep.* 2020; **10**: 5245.
31 Pando E, Alberti P, Mata R *et al*. Early changes in blood urea nitrogen (BUN) can predict mortality in acute pancreatitis: comparative study between BISAP Score, APACHE-II, and other laboratory markers-a prospective observational study. *Can. J. Gastroenterol. Hepatol.* 2021; **2021**: 6643595.
32 Chen H, Huang L, Jiang X *et al*. Establishment and analysis of a disease risk prediction model for the systemic lupus erythematosus with random forest. *Front. Immunol.* 2022; **13**.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Clinical characteristics and risk factors for patients with AP in the training and validation cohorts.

Table S2. Univariate analysis in the training cohort.

Figure S3. Standard box plots presenting the distributions of the seven significant factors for SAP. NS: non-severe acute pancreatitis; S: severe acute pancreatitis; T: training cohort; V: validation cohort.

Table S4. Multivariate analysis in the training cohort.