Application Value of the Automated Machine Learning Model Based on Modified Computed Tomography Severity Index Combined With Serological Indicators in the Early Prediction of Severe Acute Pancreatitis

Rufa Zhang, MMed, Minyue Yin, MMed,† Anqi Jiang, MMed,* Shihou Zhang, MMed,* Luojie Liu, MD,* and Xiaodan Xu, MD**

Background and Aims: Machine learning (ML) algorithms are widely applied in building models of medicine due to their powerful studying and generalizing ability. To assess the value of the Modified Computed Tomography Severity Index (MCTSI) combined with serological indicators for early prediction of severe acute pancreatitis (SAP) by automated ML (AutoML).

Patients and Methods: The clinical data, of the patients with acute pancreatitis (AP) hospitalized in Hospital 1 and hospital 2 from January 2017 to December 2021, were retrospectively analyzed. Serological indicators within 24 hours of admission were collected. MCTSI score was completed by noncontrast computed tomography within 24 hours of admission. Data from the hospital 1 were adopted for training, and data from the hospital 2 were adopted for external validation. The diagnosis of AP and SAP was based on the 2012 revised Atlanta classification of AP. Models were built using traditional logistic regression and AutoML analysis with 4 types of algorithms. The performance of models was evaluated by the receiver operating characteristic curve, the calibration curve, and the decision curve analysis based on logistic regression and decision curve analysis, feature importance, SHapley Additive exPlanation Plot, and Local Interpretable Model Agnostic Explanation based on AutoML.

Results: A total of 499 patients were used to develop the models in the training data set. An independent data set of 201 patients was used to test the models. The model developed by the Deep Neural Net (DL) outperformed other models with an area under the receiver operating characteristic curve (areas under the curve) of 0.907 in the test set. Furthermore, among these AutoML models, the DL and gradient boosting machine models achieved the highest sensitivity values, both exceeding 0.800.

Conclusion: The AutoML model based on the MCTSI score combined with serological indicators has good predictive value for SAP in the early stage.

(*J Clin Gastroenterol* 2024;58:692–701)

Acute pancreatitis (AP) is an inflammatory disorder of the pancreas involving local and peripancreatic tissue. Although AP is inclined to be self-limiting, organ failure (OF) is a hallmark complication of severe AP (SAP) and may be found in ~20% of all cases of AP.[1] The mortality rate of AP increases as much as 30% when OF occurs.[2] Therefore, it is crucial to predict the risk of SAP at an early phase, so that early treatment, such as fluid resuscitation, is dispensable for reducing the morbidity and mortality of SAP.

Conventional scoring systems, such as the Ranson's criteria (RANSON) score, bedside index of severity in AP (BISAP), Modified Computed Tomography Severity Index (MCTSI), and Acute Physiology and Chronic Health Evaluation II, have been generally applied to assess the severity of AP,[3,4] but showed modest value in early predicting possible SAP. In addition, the traditional MCTSI score often requires enhanced computed tomography (CT) to be performed, which can be difficult to achieve in the early stages and may not be suitable for patients with relevant contraindications.

Machine learning (ML) considered as a subset of artificial intelligence, was not only applied in text mining and classification in the field of computer science but also widely used in clinical practice. The novel automated ML (AutoML) intelligently selects from various algorithms and hyperparameters to create models customized to target data. It takes less time to develop more accurate models using intelligent early stopping, cross-validation, regularization, and hyperparameter optimization when compared with traditional ML, such as logistic regression (LR), support vector machine, and so on.

Our study aims to assess the value of MCTSI combined with serological indicators for early prediction of SAP within 24 hours of hospitalization, using the H2O AutoML platform in 2 hospitals. Being compared with traditional LR and 3 existing score systems.

PATIENTS AND METHODS

Inclusion and Exclusion Criteria

A retrospective analysis was performed in the 2 hospitals from January 2017 to December 2021. The data set gathered from patients from January 2017 to December 2021 was regarded as the training set in hospital 1, and the data set gathered from patients from January 2019 to December 2020

Received for publication April 18, 2023; accepted July 16, 2023.

From the *Department of Gastroenterology, Changshu Hospital Affiliated to Soochow University, Changshu No. 1 People's Hospital; and †Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China.

This study was funded by Gusu Health Personnel Training Project (GSWS2020109).

The authors declare that they have nothing to disclose.

Address correspondence to: Xiaodan Xu, MD, No. 1 Shuyuan Street, Suzhou, Jiangsu 215500, China (e-mail: xxd20@163.com).

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.jcge.com.

DOI: 10.1097/MCG.0000000000001909

was recorded as the test set. Two hospitals are large-scale and fully equipped tertiary teaching hospitals in China. There are 1320 beds in hospital 1 and more than 3000 beds in hospital 2, as a county hospital, successfully established 5 major centers, including a chest pain center, stroke center, atrial fibrillation center, etc.

The diagnostic criteria for AP were set up according to the revised Atlanta classification of AP 2012. At least two of the following 3 criteria had to be satisfied for a diagnosis of AP: (1) typical abdominal pain, (2) serum amylase beyond three times the upper limit of normal, and (3) images of characteristic findings of AP.[5] Adult patients (≥ 18 y old) who were diagnosed with AP based on the criteria were enrolled. Severe AP (SAP) was defined as AP with persistent organ failure (>48 h). Patients were divided into 2 groups: SAP and non-SAP. The exclusion criteria were patients who had chronic liver disease, chronic renal disease, hematological diseases, recurrent/chronic/traumatic/idiopathic pancreatitis, pancreatic cancer, history of pancreatic resection, patients who experienced chemoradiotherapy, and patients who were pregnant. All patients were treated in accordance with the guidelines for the management of AP. This study was approved by the ethics committee of the hospital 1 (Fig. 1).

Data Collection

Demographic characteristics and clinical information and concomitant diseases were extracted from electronic medical records. Laboratory data within 24 hours of admission were collected, including blood routine examinations, coagulation tests, and serum biochemical tests. Also, the presence of pleural effusion (PE) and MCTSI score were recorded according to the CT scan within 24 hours of admission. Finally, a total of 43 variables were extracted for analysis. Details are listed in Supplemental Table S1 (Supplemental Digital Content 1, http://links.lww.com/JCG/B2). Missing variables, which were recognized as missing data at random, were multiply imputed using a random forest algorithm by the "mice" package of R software.[6] The scoring systems, such as systemic inflammatory response syndrome (SIRS), RANSON, MCTSI, and BISAP were calculated, as described,[3,4,7] if data were available. The flowchart of this study is shown in Figure 1.

Logistic Regression

To address multicollinearity among predictor variables, the least absolute shrinkage and selection operator (LASSO) regression model was used for univariate analysis, using the "λ_1se" as the criterion. The model specification was performed through binary logistic backward stepwise regression analysis. The predictive performance of the proposed model was assessed using the receiver operating characteristic (ROC) curve, the calibration curve, and the decision curve analysis (DCA). A nomogram was created based on the independent risk factors determined in the multivariate analysis.

Automated Machine Learning

The H2O package installed from the H2O.ai platform (www.h2o.ai) was applied to implement AutoML analysis, which automatically selects applicable algorithms and integrates them into multiple ensemble models. Algorithms include a default random forest (DRF), a random grid of gradient boosting machines (GBMs), an extremely randomized forest, a random grid of deep neural nets (DLs), and a fixed grid of generalized linear models (GLMs). The training set underwent a 5-fold cross-validation grid search to optimize hyperparameters, and the effectiveness of various hyperparameter combinations was verified by assessing the areas under the curve (AUCs). The confusion matrix, consisting of true positives (TPs), true negatives (TN), false positives (FP), and false negatives (FN), was established to calculate sensitivity, specificity, positive predictive value, negative predictive value (NPV), positive likelihood ration (LR+), negative likelihood ration (LR−), accuracy (ACC), and areas under the ROC curve (AUCs) for evaluating discrimination performance of models. Formulas were as follows: ACC = (TP+TN)/(TP+FP+FN+TN); positive predictive value = TP/(TP+NP); NPV = TN/(TN+FN); LR+ = sensitivity/(1-specificity); LR− = (1-sensitivity)/specificity. The visualization of AutoML was exhibited in the form of feature importance, SHapley Additive exPlanation (SHAP), and local interpretable model agnostic explanation (LIME). SHAP analysis explained which features were most important for creating model predictions and how much they contributed to the overall model performance for a particular prediction.[8] The LIME analysis demonstrated how much each feature contributed to predicting the outcome by randomly giving examples from the test set.

Statistical Analyses

Continuous variables were expressed as mean ± SD if fitting a normal distribution and as median (interquartile range) if not. Categorical variables were shown as frequencies. We compared the two groups by the Pearson χ^2 test or Fisher exact tests for categorical variables and the Student *t* test or nonparametric Mann-Whitney *U* test for continuous variables. A 2-sided *P* value of <0.05 was considered statistically significant. Analyses were performed with R software (version 4.2.1), including the H2O package (version 3.36.0.2), tableone package (version 0.12.0), tidyverse package (version 1.3.0), tidyquant package (version 1.0.2), and LIME package (version0.5.1).

RESULTS

Baseline Characteristics

A total of 700 patients with AP were included in our study. SAP occurred in 63 cases (9.0%) in the whole cohort. In these SAP cases, the occurrence of local complications is as follows: 56 cases (88.9%) of acute peripancreatic fluid collection, 14 cases (22.2%) of acute necrotic collection, 6 cases (9.5%) of walled-off necrosis, and 11 cases (17.5%) of pancreatic pseudocyst, and the occurrence of systemic complications is as follows: 40 cases (63.5%) of SIRS, 56 cases (88.9%) of respiratory dysfunction, 14 cases (22.2%) of circulatory dysfunction, 10 cases (15.9%) of renal dysfunction, 13 cases (20.6%) of multiple organ dysfunction syndrome, 2 cases (3.2%) of death, and 1 case (1.6%) with unknown outcome. Among these patients, 499 patients from hospital 1 were included in the training data set. A total of 201 patients from the hospital 2 were selected as a test data set. In the training data set, 56.9% (284/499) were men and 43.1% (215/499) were women. The median age was 53 years, ranging from 41 to 66 in the non-SAP group and 42 ranging from 34 to 53.5 in the SAP group. In the test data set, the onset of AP and SAP were also more commonly seen in male than in female patients and the median age ranged from 48 to 48.5 years. Consistent with what was reported,[1] biliary sludge or gallstones (39.49%) was the most frequent etiology of AP in our cohorts, followed by hypertriglyceridemia (17.87%). No statistical differences were observed in sex, smoke, history of hypertension, and diabetes in two groups of 3 data sets ($P > 0.05$). Details are listed in Table 1.

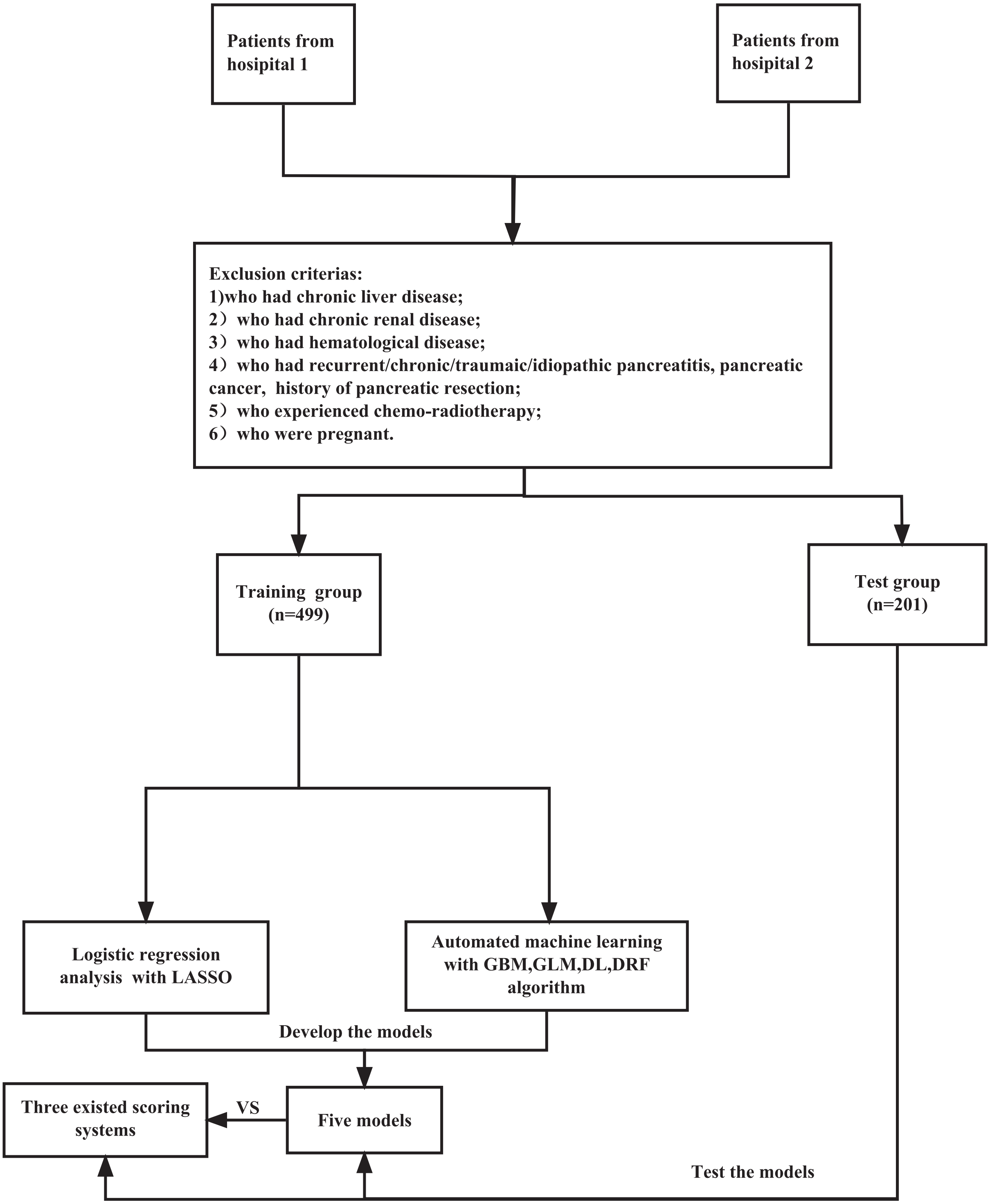
Patients from hosipital 1
Patients from hosipital 2
Exclusion criterias:
1)who had chronic liver disease;
2）who had chronic renal disease;
3）who had hematological disease;
4）who had recurrent/chronic/traumaic/idiopathic pancreatitis, pancreatic cancer, history of pancreatic resection;
5）who experienced chemo-radiotherapy;
6）who were pregnant.
Training group (n=499)
Test group (n=201)
Logistic regression analysis with LASSO
Automated machine learning with GBM,GLM,DL,DRF algorithm
Develop the models
Three existed scoring systems
VS
Five models
Test the models

FIGURE 1. The flowchart of the study. DL indicates deep neural net DRF, default random forest; GBM, gradient boosting machine; GLM, generalized linear model; LASSO, least absolute shrinkage and selection operator.

Univariate and Multivariate Logistic Regression Analysis

Three variables of the 43 variables were selected and later reserved as independent risk factors using the LASSO regression model with the "λ_1se (0.035)" criterion, which was achieved by 5-fold cross-validation, to solve such multiple colinear relationships among the explanatory variables (Fig. 2). The final logistic model, including 3 variables [white blood cell (WBC), Ca^{2+} and MCTSI], was developed as a nomogram and a score system for clinical use (Fig. 3). The calibration curves of the training and test set are plotted in Figure 4, and the mean absolute errors were both 0.023,

TABLE 1. Baseline Characteristics of Patients in Training and Test Groups

		Training data set (n = 499)			Test data set (n = 201)		
Variables	**Group**	**Non-SAP (n = 452); n (%)**	**SAP (n = 47); n (%)**	***P***	**Non-SAP (n = 185); n (%)**	**SAP (n = 16); n (%)**	***P***
Sex	Male	253 (56.0)	31 (66.0)	0.246	111 (60.0)	9 (56.2)	0.978
	Female	199 (44.0)	16 (34.0)	—	74 (40.0)	7 (43.8)	—
Age (y); median (IQR)	—	53.00 (41.00, 66.00)	42.00 (34.00, 53.50)	< 0.001	48.00 (36.00, 61.00)	48.50 (34.25, 68.00)	0.717
Etiology	Biliary	204 (45.1)	9 (19.1)	< 0.001	68 (36.8)	9 (56.2)	0.217
	Hyperlipidemia	71 (15.7)	26 (55.3)	—	52 (28.1)	5 (31.2)	—
	Alcoholic	25 (5.5)	1 (2.1)	—	10 (5.4)	1 (6.2)	—
	Others	152 (33.6)	11 (23.4)	—	55 (29.7)	1 (6.2)	—
Smoke	No	409 (90.5)	39 (83.0)	0.233	140 (75.7)	14 (87.5)	0.445
	Yes	42 (9.3)	8 (17.0)	—	45 (24.3)	2 (12.5)	—
Hypertension	No	302 (66.8)	32 (68.1)	0.989	125 (67.6)	8 (50.0)	0.25
	Yes	150 (33.2)	15 (31.9)	—	60 (32.4)	8 (50.0)	—
Diabetes	No	409 (90.5)	43 (91.5)	1	147 (79.5)	11 (68.8)	0.494
	Yes	43 (9.5)	4 (8.5)	—	38 (20.5)	5 (31.2)	—
SBP; mean (SD)	—	133.57 (18.75)	132.43 (22.77)	0.698	132.78 (15.32)	131.50 (26.14)	0.764
DBP; mean (SD)	—	79.94 (11.99)	82.21 (15.61)	0.232	79.11 (10.40)	79.44 (12.45)	0.906
MAP; median (IQR)	—	96.67 (90.00, 104.75)	96.67 (89.83, 105.17)	0.883	96.67 (89.33, 104.33)	96.66 (90.17, 104.58)	0.761
PLT ($\times 10^9$/L); median (IQR)	—	181.50 (146.00, 228.00)	209.00 (155.50, 271.00)	0.035	207.00 (171.00, 242.00)	211.50 (186.25, 286.50)	0.302
WBC ($\times 10^9$/L); median (IQR)	—	11.10 (8.70, 14.70]	16.40 (12.00, 20.75)	< 0.001	13.30 (10.07, 15.86)	18.27 (14.36, 21.20)	0.001
Neutrophil count ($\times 10^9$/L); median (IQR)	—	9.10 (6.38, 12.10)	13.90 (10.20, 17.55)	< 0.001	11.08 (7.95, 13.55)	16.42 (12.18, 19.42)	< 0.001
Lymphocyte count ($\times 10^9$/L); median (IQR)	—	1.30 (0.90, 1.90)	1.20 (0.95, 1.90)	0.78	1.16 (0.79, 1.83)	0.92 (0.74, 1.10)	0.025
NLR; median (IQR)	—	6.98 (3.95, 11.81)	10.83 (6.14, 15.27)	0.003	8.63 (4.73, 15.96)	17.59 (12.92, 23.00)	< 0.001
HCT (L/L); median (IQR)	—	0.42 (0.39, 0.46)	0.45 (0.42, 0.48)	0.006	0.42 (0.39, 0.46)	0.41 (0.37, 0.44)	0.236
RDW (%); median (IQR)	—	12.80 (12.30, 13.40)	12.80 (12.50, 13.30)	0.636	12.80 (12.40, 13.20)	12.90 (12.47, 13.62)	0.313
Lr (%); median (IQR)	—	11.95 (7.40, 18.60)	8.00 (5.75, 12.90)	0.002	9.80 (5.60, 16.10)	5.15 (3.98, 6.92)	< 0.001
Cr (μmol/L); median (IQR)	—	66.00 (55.00, 77.00)	60.00 (52.00, 76.00)	0.218	61.40 (50.90, 72.40)	59.15 (55.12, 87.12)	0.229
TB (μmol/L); median (IQR)	—	21.35 (14.57, 35.02)	19.10 (13.65, 30.70)	0.174	18.60 (14.20, 24.50)	16.75 (12.73, 27.50)	0.522
DB (μmol/L); median (IQR)	—	7.50 (4.50, 15.00)	6.30 (3.50, 14.15)	0.159	6.00 (4.10, 8.30)	6.15 (3.10, 9.82)	0.882
DTR; median (IQR)	—	0.37 (0.29, 0.52)	0.41 (0.26, 0.52)	0.89	0.31 (0.26, 0.39)	0.32 (0.26, 0.44)	0.656
Urea (mmol/L); median (IQR)	—	4.80 (3.80, 6.00)	4.90 (4.00, 6.15)	0.788	4.80 (3.90, 6.00)	7.05 (5.07, 9.68)	0.002
LDH (U/L); median (IQR)	—	231.50 (184.00, 342.25)	315.00 (223.00, 399.50)	0.002	205.10 (176.10, 246.00)	408.15 (232.33, 545.38)	< 0.001
Ca^{2+}(mmol/L); median (IQR)	—	2.26 (2.16, 2.36)	1.90 (1.74, 2.11)	< 0.001	2.10 (2.03, 2.18)	1.92 (1.75, 2.04)	< 0.001
TG (mmol/L); median (IQR)	—	1.42 (0.88, 3.12)	9.43 (1.65, 22.20)	< 0.001	1.57 (0.93, 3.75)	2.10 (1.32, 6.41)	0.231
Glucose (mmol/L); median (IQR)	—	7.40 (6.25, 9.59)	8.13 (6.20, 11.09)	0.289	6.98 (5.68, 8.58)	9.35 (7.21, 12.22)	0.003
TyG; median (IQR)	—	9.06 (8.47, 9.82)	10.96 (9.43, 11.67)	< 0.001	9.05 (8.38, 10.11)	9.56 (8.96, 10.85)	0.057
ALT (U/L); median (IQR)	—	49.50 (22.00, 175.50)	29.00 (13.50, 59.50)	0.002	26.80 (15.10, 99.90)	29.85 (17.65, 88.60)	0.638
AST (U/L); median (IQR)	—	41.00 (22.00, 161.00)	27.00 (21.00, 56.50)	0.029	22.00 (16.00, 52.90)	28.10 (15.50, 59.80)	0.544
GGT (U/L); median (IQR)	—	105.50 (37.75, 310.25)	76.00 (33.50, 248.00)	0.607	60.50 (30.00, 147.50)	66.75 (34.77, 222.35)	0.548
ALP (U/L); median (IQR)	—	104.50 (79.75, 157.00)	88.00 (72.50, 131.00)	0.092	66.90 (55.30, 93.10)	63.30 (57.67, 86.80)	0.749
Albumin (g/L); mean (SD)	—	38.50 (5.06)	38.00 (6.24)	0.535	36.85 (4.57)	33.76 (7.19)	0.015
K^{+}(U/L); median (IQR)	—	3.92 (3.67, 4.19)	3.83 (3.62, 4.02)	0.179	4.17 (3.87, 4.44)	4.02 (3.88, 4.78)	0.649
AGR; median (IQR)	—	1.41 (1.21, 1.61)	1.23 (1.04, 1.35)	< 0.001	1.30 (1.20, 1.60)	1.15 (1.00, 1.42)	0.038
PT (s); median (IQR)	—	13.20 (12.30, 14.33)	13.90 (12.25, 15.85)	0.061	12.60 (11.90, 13.50)	14.30 (13.35, 14.60)	< 0.001
INR; median (IQR)	—	1.06 (0.99, 1.16)	1.13 (1.02, 1.30)	0.025	1.06 (1.02, 1.13)	1.15 (1.10, 1.21)	0.002
APTT (s); median (IQR)	—	32.55 (29.28, 36.62)	32.40 (29.70, 36.50)	0.9	29.50 (25.50, 33.90)	35.60 (31.65, 40.85)	0.001

TABLE 1. (*continued*)

		Training data set (n = 499)			Test data set (n = 201)		
Variables	**Group**	**Non-SAP (n = 452); n (%)**	**SAP (n = 47); n (%)**	***P***	**Non-SAP (n = 185); n (%)**	**SAP (n = 16); n (%)**	***P***
CRP (mg/L); median (IQR)	—	9.00 (2.10, 53.40)	29.80 (6.50, 161.80)	< 0.001	86.90 (16.40, 194.31)	190.36 (84.58, 293.49)	0.023
CAR; median (IQR)	—	0.22 (0.05, 1.39)	0.72 (0.17, 3.99)	0.001	2.46 (0.33, 5.16)	6.14 (1.34, 8.75)	0.044
RCR; median (IQR)	—	5.64 (5.34, 6.12)	6.99 (5.79, 7.58)	< 0.001	6.12 (5.81, 6.41)	6.78 (6.20, 7.97)	< 0.001
SIRS	No	356 (78.8)	22 (46.8)	< 0.001	117 (63.2)	1 (6.2)	< 0.001
	Yes	96 (21.2)	25 (53.2)	—	68 (36.8)	15 (93.8)	—
PE	No	308 (68.1)	8 (17.0)	< 0.001	139 (75.1)	1 (6.2)	< 0.001
	Yes	143 (31.6)	39 (83.0)	—	46 (24.9)	15 (93.8)	—
MCTSI; median (IQR)	—	2.00 (2.00, 4.00)	4.00 (4.00, 4.00)	< 0.001	2.00 (2.00, 2.00)	4.00 (4.00, 4.00)	< 0.001
RANSON; median (IQR)	—	1.00 (0.00, 2.00)	1.00 (1.00, 2.00)	0.095	1.00 (0.00, 2.00)	2.00 (1.00, 2.00)	< 0.001
BISAP; median (IQR)	—	1.00 (0.00, 1.00)	2.00 (1.00, 2.00)	< 0.001	1.00 (0.00, 2.00)	2.00 (2.00, 3.00)	< 0.001

AGR indicates albumin/globulin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; APTT, activated partial thromboplastin time; AST, aspartate aminotransferase; BISAP, bedside index of severity in acute pancreatitis; Cr, creatinine; DB, direct bilirubin; DBP, diastolic blood pressure; DTR, direct total bilirubin; GGT, gamma-glutamyl transferase; HCT, hematocrit; IQR, interquartile range; INR, international normalization ratio; LDH, lactate dehydrogenase; Lr, percentage of lymphocytes; MAP, mean artery pressure; MCTSI, Modified Computed Tomography Severity Index; NLR, neutrophil lymphocyte count; PE, pleural effusion; PLT, platelet count; PT, prothrombin time; RANSON, Ranson's criteria; RCR, ratio of red cell distribution width (RDW) to Ca2+; RDW, red cell distribution width; SAP, severe acute pancreatitis; SBP, systolic blood pressure; SIRS, systemic inflammatory response syndrome; TB, total bilirubin; TG, total triglycerides; TyG, triglycerides to glucose; WBC, white blood cell.

respectively, demonstrating that the estimated risk using the LASSO model was close to the observed risk, indicating a high degree of reliability. The DCA plots of the training set and test set are presented in Figure 5, demonstrating that when the threshold probability of SAP predicted by the LASSO model was between 10% and 90%, an intervention might add more benefit (1% to 7%). When a clinician considered the patient had a 20% chance of developing SAP, the patient might gain 4% of the benefit from early intervention, according to the DCA of the test set, which is equivalent to detecting 4 patients with SAP and suggesting zero unnecessary treatment per 100 patients. This is a direct comparison with treat none (the horizontal line in Fig. 5), which has zero TPs and zero FPs by default.[9] The net benefit declares that the use of the LASSO model would improve patient outcomes irrespective of patient or doctor preference. The ROC curve of the test set is presented in Supplemental Figure S1 (Supplemental Digital Content 1, http://links.lww.com/JCG/B2), and its AUC was 0.906 as shown in Table 2.

Automated Machine Learning Analysis

A total of 62 models were developed based on 4 ML algorithms (DL, GBM, GLM, and DRF), and stacked ensemble models were removed because of poor interpretability. The GBM and DRF models were the best among these models due to their highest AUC of 1.000, which was a comprehensive evaluation for imbalanced samples. As shown in Figure 6, Ca^{2+} was the most important feature, followed by total triglycerides (TGs), the ratio of red cell distribution width (RDW) to Ca^{2+} (RCR), MCTSI, WBC, PE, the ratio of RDW, C-reactive protein (CRP), triglycerides to glucose (TyG) ratio, and international normalization ratio in the GBM model. In addition, Ca^{2+}, WBC, and MCTSI were the important variables in common between the GBM model and the LASSO model. SHAP contribution plots based on GBM algorithms are presented in Figure 7, including 10 important variables [Ca^{2+}, PE, ratio of RDW to Ca^{2+} (RCR), MCTSI, neutrophil count, TGs, CRP, K^+, C-reactive protein/albumin (CAR), and creatinine]. The closer the values of the variables were to 1, the more likely patients were to progress to SAP. For example, the red part of PE, which was concentrated on the right of axis = 0, revealed that the patients with AP with PE would be more likely to develop SAP. Conversely, the red part of Ca^{2+} was concentrated on the left of axis = 0, which suggests that for patients with AP, an increase in serum Ca^{2+} levels may lead to a decreased likelihood of developing SAP. Table 2 demonstrates that the DL algorithm achieved a higher value of AUC than the GBM, GLM, and DRF algorithms (0.907, 0.895, 0.855, and 0.846, respectively). The accuracies were 0.915, 0.910, 0.896, and 0.851 according to the confusion matrix of GLM, DL, GBM, and DRF, respectively, models on the test set. A LIME plot of the GBM model on the test set exhibited how several important variables contributed to the progress of SAP. As shown in Figure 8, for example, case 1 had a high probability of 0.93 for progressing to SAP as predicted by the GBM model. Ca^{2+} was the most significant feature contributing to the prediction, followed by PE, whereas RCR, TG, and CRP had the opposite effect. The DCA plots of the test set are presented in Supplemental Figure S2 (Supplemental Digital Content 1, http://links.lww.com/JCG/B2), demonstrating that when the threshold probability of SAP predicted by the AutoML models was between 10% and 100%, an intervention might add more benefit (1% to 7%). When a clinician considered the patient had a 10% chance of developing SAP, the patient

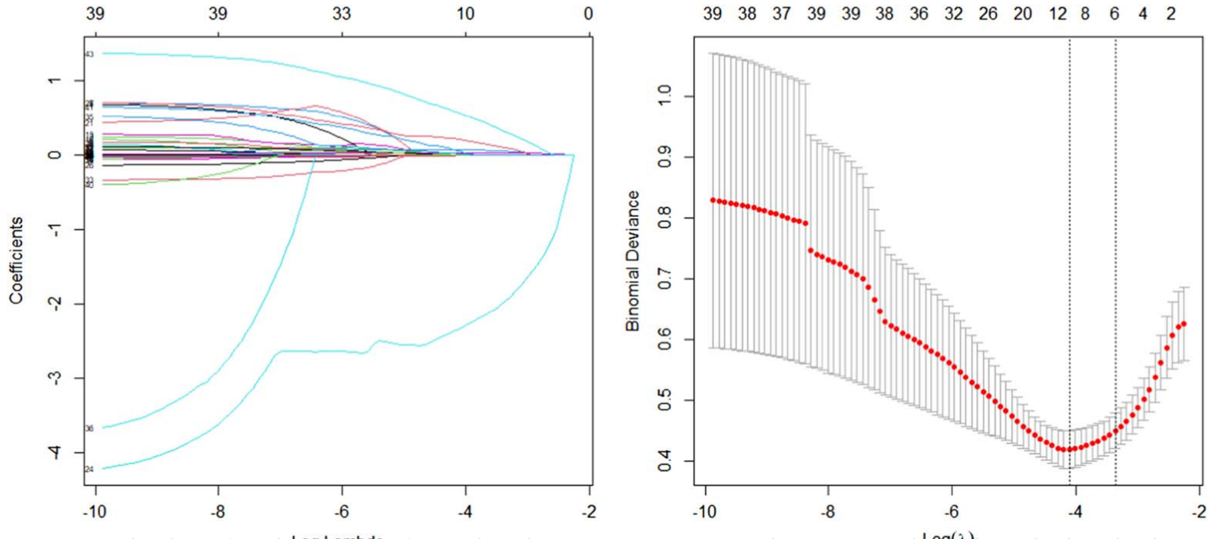

FIGURE 2. Penalty chart of predictive factors for SAP based on LASSO regression analysis. LASSO indicates least absolute shrinkage and selection operator; SAP, severe acute pancreatitis.

might gain 5% of the benefit from early intervention, according to the DCA of the test set, which is equivalent to detecting 5 patients with SAP and suggesting zero unnecessary treatment per 100 patients.

Comparisons Between Existing Scoring Systems and Models Developed by Logistic Regression and Automated Machine Learning

In the test set, the AUC values obtained by the eight models were 0.907 for DL, 0.906 for LASSO, 0.895 for GBM, 0.855 for GLM, 0.849 for BISAP, 0.846 for DRF, 0.812 for MCTSI, and 0.790 for RANSON. The RANSON, BISAP, MCTSI, DL, and GLM models achieved the highest accuracy among these models, beyond 0.900. MCTSI achieved the highest specificity value of 0.990 and the lowest sensitivity of 0.125. The LASSO achieved a high sensitivity value of 0.938 and the lowest specificity of 0.795. Details are listed in Table 2.

DISCUSSION

In this study, we developed and tested several AutoML models to early identify who would progress to SAP. These models were almost superior to existing scoring systems such as BISAP, RANSON, and MCTSI. In addition, the DL model obtained the highest value of AUC above 0.900, with specificity and accuracy all above 0.910. Early prediction of patients with SAP is essential for determining which patients require appropriate management, such as intensive care, rapid liquid resuscitation, and early enteral nutrition.[10] Up to now, various scoring systems have been developed for early risk stratification of patients with AP. Some novel point systems, such as acute biliary pancreatitis point systems,[7] the pancreatic activity scoring system, and the Chinese simple scoring system,[11] have been proposed in recent years. Typical models, such as RANSON, MCTSI, and BISAP, in our study, achieved inferior accuracy to the models we built. And, the traditional scores are relatively complicated for clinical use, and the novel scores are not generalized, whose ability to predict SAP varies and accuracy ranges from 0.70 to 0.95.[7,11,12] In addition, our aim is to early predictions of SAP, making sensitivity crucial for

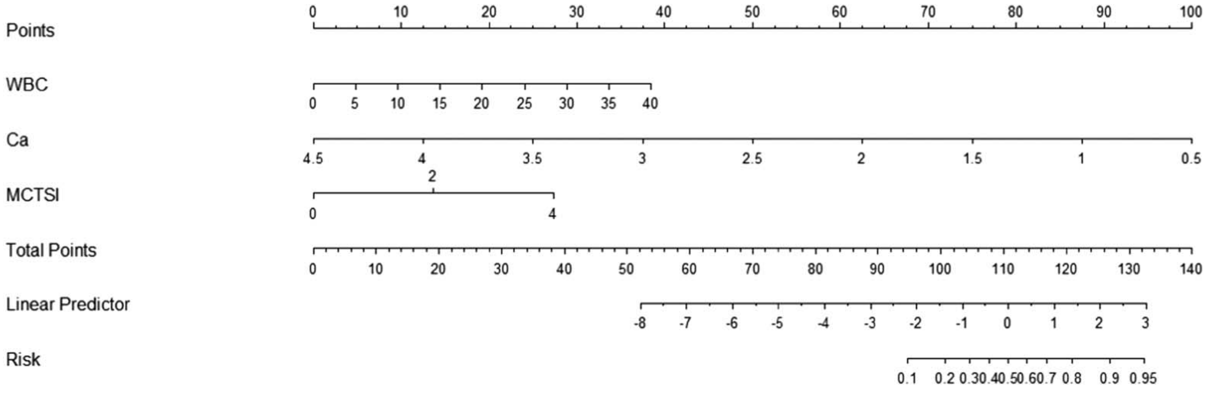

FIGURE 3. Nomogram of the LASSO model for the early prediction of SAP. LASSO indicates least absolute shrinkage and selection operator; MCTSI, Modified Computed Tomography Severity Index; SAP, severe acute pancreatitis; WBC, white blood cell.

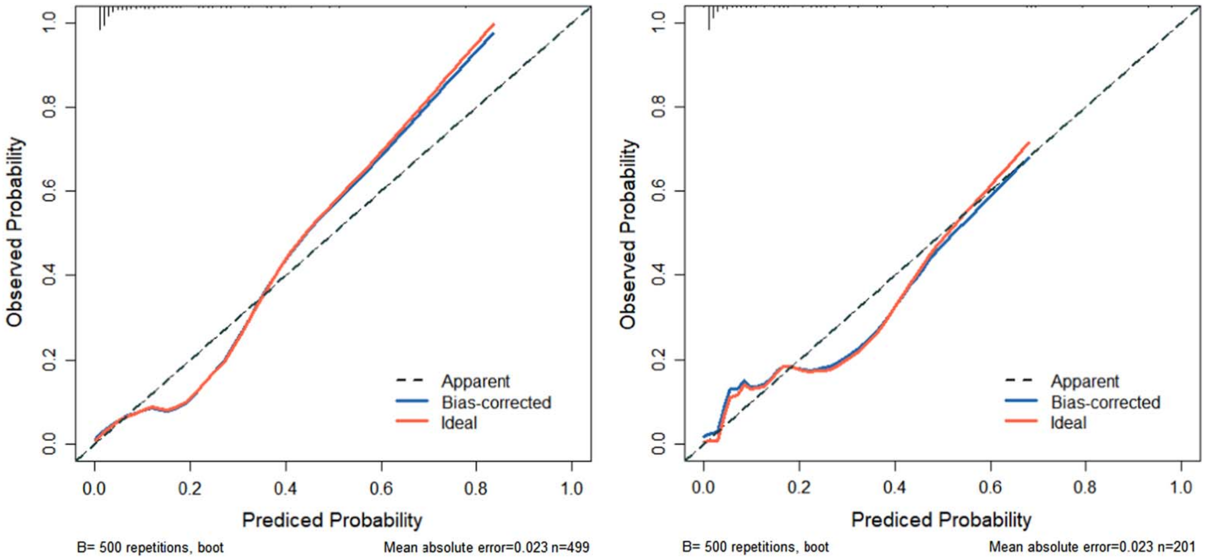

FIGURE 4. Calibration curve of the LASSO model in the training and test set. LASSO indicates least absolute shrinkage and selection operator.

our purposes. Although the MCTSI has high specificity, it is sensitivity is relatively low. In contrast, by combining serological indicators, we can substantially improve sensitivity without a noticeable change in specificity.

Compared with traditional univariate and sequent multivariate analyses, AutoML greatly improved work efficiency due to its less time consumption and higher accuracy. In addition, ensemble models combined various ML algorithms, utilizing multiclassifiers to predict the target outcome by taking a vote of individual predictions, which could enhance the overall performance.[13] In this study, we selected 4 models built by 4 types of AutoML algorithms (GBM, DRF, GLM, and DL) for predicting the risk of SAP. All models, among which the DL model ranked first in AUC, DL and GLM in accuracy on the test data set yielded satisfactory results. AUC gives a more feasible method to settle the problem of unbalanced data by putting the same weight on both classes in contrast to accuracy.[14]

The SHAP analysis demonstrated that the occurrence of Ca^{2+} at admission was the most important feature of the GBM model. In our study, Ca^{2+} was a common important feature selected not only by GBM but also by LASSO and RANSON, indicating that Ca^{2+} was indeed a reliable serological indicator in the prediction of SAP. Peng et al[15] reported a significant correlation between the decrease of serum Ca^{2+} and the incidence of persistent organ failure by triggering the SIRS process that recruits neutrophils and leads to further release of reactive oxygen species and organ damage. Chen et al[16] carried out a subanalysis in hypertriglyceridemia pancreatitis populations for exploring the association between albumin and the severity of AP. It was generally believed that elevated levels of TG would drive the

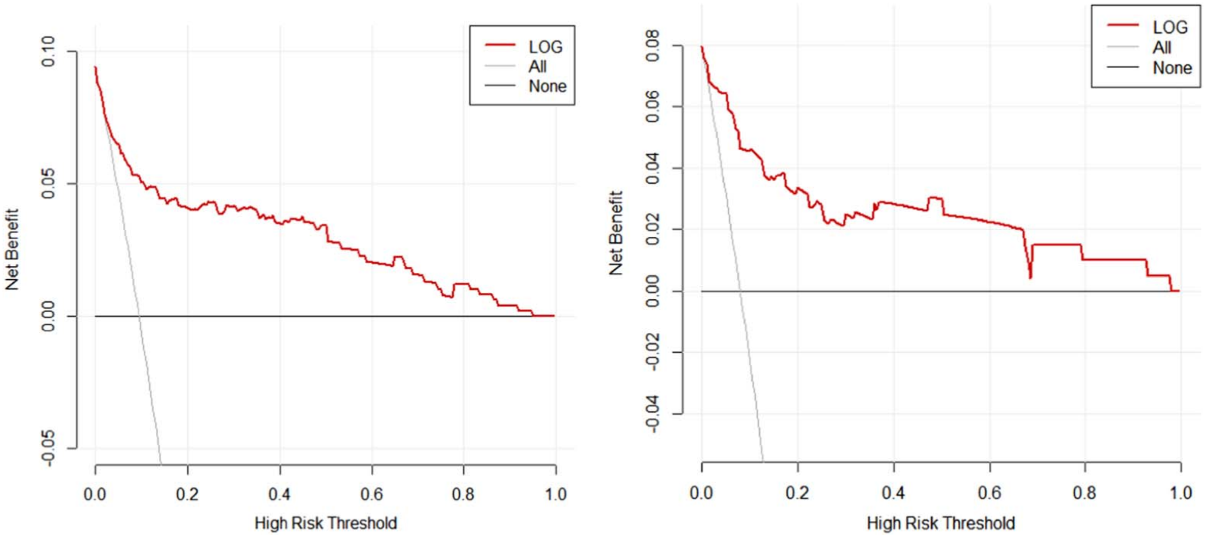

FIGURE 5. Decision curve analysis of the LASSO model in the training and test set. LASSO indicates least absolute shrinkage and selection operator.

TABLE 2. Comparison of LR and AutoML Models for Early Prediction of SAP in the Test Cohort

AUC		Sensitivity	Specificity	Accuracy	PPV	NPV	LR+	LR−
AutoML								
GBM	0.895	0.812	0.903	0.896	0.419	0.982	8.351	0.208
DRF	0.846	0.688	0.865	0.851	0.306	0.970	5.088	0.361
GLM	0.855	0.750	0.930	0.915	0.480	0.977	10.673	0.269
DL	0.907	0.812	0.919	0.910	0.464	0.983	10.021	0.204
LR								
LASSO	0.906	0.938	0.795	0.806	0.283	0.993	4.564	0.079
Existed scoring systems								
RANSON	0.790	0.125	0.989	0.920	0.500	0.929	11.363	0.885
MCTSI	0.812	0.125	0.990	0.925	0.667	0.929	12.500	0.883
BISAP	0.849	0.125	0.989	0.920	0.500	0.929	11.363	0.885

AUC indicates area under the curve; AutoML, automated machine learning; BISAP, bedside index of severity in acute pancreatitis; DL, deep neural net; DRF, default random forest; GBM, gradient boosting machine; LASSO, least absolute shrinkage and selection operator; LR, logistic regression; LR−, negative likelihood ratio; LR+, positive likelihood ratio; MCTSI, Modified Computed Tomography Severity Index; NPV, negative predictive value; PPV, positive predictive value; RANSON, Ranson's criteria; SAP, severe acute pancreatitis.

occurrence of SAP due to toxic effects on pancreatic acinar cells.[16] The free fatty acids, hydrolyzed by pancreatic lipase from TG, can bind to albumin in the serum and thus stimulate the inflammatory process. Therefore, Chen and colleagues' study effectively ruled out the confounding effect of TG and demonstrated that the decrease in albumin was indeed an independent predictive factor. Both in GBM, DRF and GLM, TG is an important feature for predictive models.

Another study proposed that RDW, a marker reelecting inflammation status, showed great predictive performance of AP severity with an AUC of >0.810 and mortality with an AUC of >0.842.[17] In addition, this study further suggested that the ratio of RDW to Ca^{2+} (RCR) was an excellent predictor of AP severity with an AUC value of 0.973. Han et al[18] also discovered a positive correlation between a high level of RCR and a poor prognosis for patients with AP. Consistent with the aforesaid studies, our study illustrated that Ca^{2+}, RCR, and RDW were among the top 10 important variables in the GBM model.

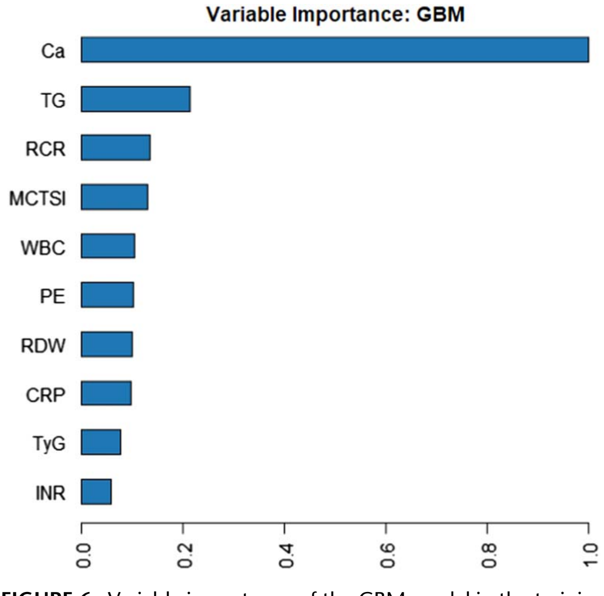

FIGURE 6. Variable importance of the GBM model in the training set. CRP indicates C-reactive protein; GBM, gradient boosting machine; INR, international normalization ratio; MCTSI, Modified Computed Tomography Severity Index; PE, pleural effusion; RCR, ratio of red cell distribution width (RDW) to Ca^{2}+; RDW, red cell distribution width; TG, total triglycerides; TyG, triglycerides to glucose; WBC, white blood cell.

The inflammatory response can cause an increase in the WBC count. Studies have shown that the WBC level 24 hours after admission can predict the severity of both gallstone and hypertriglyceridemia-induced AP.[19] And, PE is not only an important factor in the traditional RANSON scoring system but also in the novel acute biliary pancreatitis point systems scoring system for grading the severity of AP,[7,11] this finding aligns with the results of our study as well. CRP is produced by the liver in response to interleukin-6 stimulation, it typically reaches its peak within 24 to 48 hours after inflammation.[20] Numerous studies have demonstrated its close association with the severity and prognosis of AP.[21,22]

Gurda-Duda et al[23] suggested that blood glucose concentration (within 36 h from disease onset) could be a complementary measurement, with a sensitivity of 72.7% and a specificity of 75.8%. Park et al[24] investigated the association between the TyG index {= ln [fasting TG (mg/dL) × fasting plasma glucose (mg/dL)]/2} and the severity of AP in 373 patients. The results showed that the TyG index not only accurately predicted SAP but also increased the predictive value of traditional models. The underlying mechanism might be explained by insulin resistance, which activated proinflammatory molecules accelerating the progression of SAP. Several studies have demonstrated that microvessel changes are significant events in the progression of SAP and that coagulative disorders are related to SAP severity.[25,26] International normalization ratio is an indicator used to evaluate a patient's coagulation function, and it is the 10th important variable in our GBM model. However, a multivariate LR analysis performed by Radenkovic et al[27] and 3 ML algorithms performed by Qiu et al[28] did not include this parameter in the final models. Balthazar[29] proposed the CT severity index (CTSI) to assess the severity of AP, which has been used for over 20 years and is increasingly showing its limitations in the assessment of SAP. In 2004, Mortele et al[30] improved the existing CTSI system to create the MCTSI, which is more practical than the CTSI. In Mortele and colleagues' preliminary study, the MCTSI showed a better correlation with the development of

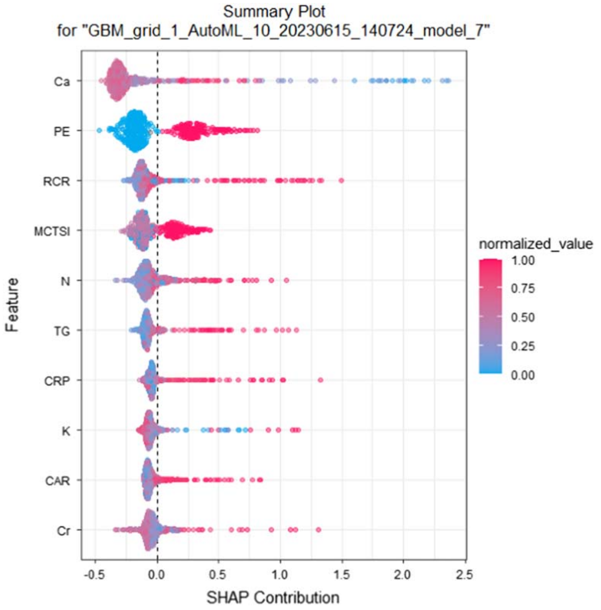

FIGURE 7. SHAP of the GBM model in the training set. CAR indicates C-reactive protein/albumin(CAR); Cr, creatinine; CRP, C-reactive protein; GBM, gradient boosting machine; MCTSI, Modified Computed Tomography Severity Index; PE, pleural effusion; RCR, ratio of red cell distribution width (RDW) to Ca^{2}+; SHAP, SHapley Additive exPlanation; TG, total triglyceride.

organ failure and length of hospital stay compared with the CTSI. Furthermore, compared with the Acute Physiology and Chronic Health Evaluation-II system, both CT scoring systems were more accurate in their correlation with pancreatic infection and intervention requirements and had higher diagnostic accuracy for clinically severe diseases.

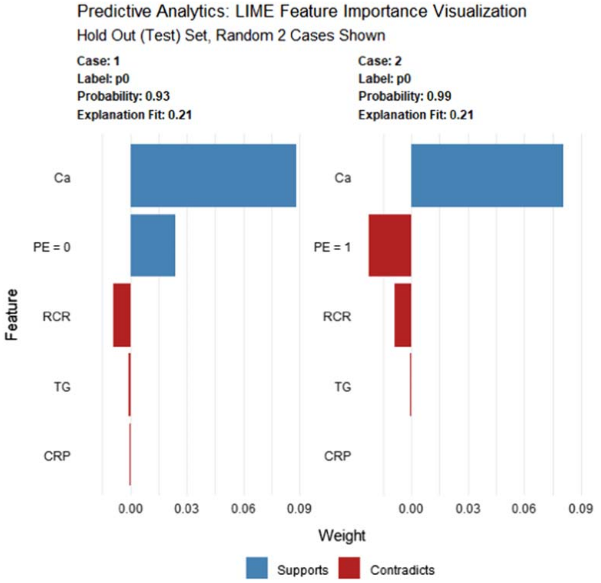

FIGURE 8. LIMEof the GBM model in the test set. CRP indicates C-reactive protein; GBM, gradient boosting machine; LIME, local interpretable model agnostic explanation; PE, pleural effusion; RCR, ratio of red cell distribution width (RDW) to Ca^{2}+; TG, total triglyceride.

This study aims to investigate the value of combining the MCTSI with multiple important serum markers in predicting SAP early.

Here, we built 5 predictive models, using traditional LR and AutoML, with a high AUC of > 0.846 and high accuracy of > 0.806. Furthermore, it is more convenient and efficient to get the predictive probability for SAP using AutoML. In addition, we used external validation in our studies, which is a common way of efficiently evaluating a new technique and may provide a better foundation for the subsequent generalization of our models. However, there are some limitations to our study. Firstly, we divided patients with AP into non-SAP and SAP instead of mild AP, moderate SAP, and SAP, which might decrease the sensitivity of our models. Secondly, our study is a retrospective study which might affect the performance of our models in a prospective clinical study. More prospective research needs to be conducted for external validation of our models. Thirdly, for our small-sample, two-center study, more multicenter and large-sample studies are needed in the future to further confirm its predictive value.

CONCLUSIONS

We developed a series of effective models for early prediction of SAP based on the AutoML platform, and these models outperformed the existing scoring systems, which might offer insights into AutoML applications in future medical studies. In addition, the DL model demonstrated practicable performance in early prediction better than LR and existing scoring systems.

REFERENCES

1. Xu F, Chen X, Li C, et al. Prediction of multiple organ failure complicated by moderately severe or severe acute pancreatitis based on machine learning: a multicenter cohort study. *Mediators Inflamm*. 2021;2021:5525118.
2. Schepers NJ, Bakker OJ, Besselink MG, et al. Impact of characteristics of organ failure and infected necrosis on mortality in necrotising pancreatitis. *Gut*. 2019;68:1044–1051.
3. Bollen TL, Singh VK, Maurer R, et al. A comparative evaluation of radiologic and clinical scoring systems in the early prediction of severity in acute pancreatitis. *Am J Gastroenterol*. 2012;107: 612–619.
4. 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;142:1476–1482; quiz e1415-1476.
5. 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;62:102–111.
6. Blazek K, van Zwieten A, Saglimbene V, et al. A practical guide to multiple imputation of missing data in nephrology. *Kidney Int*. 2021;99:68–74.
7. Hong W, Lillemoe KD, Pan S, et al. Development and validation of a risk prediction score for severe acute pancreatitis. *J Transl Med*. 2019;17:146.
8. Bang CS, Ahn JY, Kim JH, et al. Establishing machine learning models to predict curative resection in early gastric cancer with undifferentiated histology: development and usability study. *J Med Internet Res*. 2021;23:e25053.
9. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74:796–804.
10. Gliem N, Ammer-Herrmenau C, Ellenrieder V, et al. Management of severe acute pancreatitis: an update. *Digestion*. 2021;102: 503–507.

11. Wu Q, Wang J, Qin M, et al. Accuracy of conventional and novel scoring systems in predicting severity and outcomes of acute pancreatitis: a retrospective study. *Lipids Health Dis*. 2021;20:41.
12. Paragomi P, Hinton A, Pothoulakis I, et al. The modified pancreatitis activity scoring system shows distinct trajectories in acute pancreatitis: an international study. *Clin Gastroenterol Hepatol*. 2022;20:1334–1342e1334.
13. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun*. 2021;12:711.
14. Janitza S, Strobl C, Boulesteix AL. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*. 2013;14:119.
15. Peng T, Peng X, Huang M, et al. Serum calcium as an indicator of persistent organ failure in acute pancreatitis. *Am J Emerg Med*. 2017;35:978–982.
16. Chen L, Huang Y, Yu H, et al. The association of parameters of body composition and laboratory markers with the severity of hypertriglyceridemia-induced pancreatitis. *Lipids Health Dis*. 2021;20:9.
17. Gravito-Soares M, Gravito-Soares E, Gomes D, et al. Red cell distribution width and red cell distribution width to total serum calcium ratio as major predictors of severity and mortality in acute pancreatitis. *BMC Gastroenterol*. 2018;18:108.
18. Han TY, Cheng T, Liu BF, et al. Evaluation of the prognostic value of red cell distribution width to total serum calcium ratio in patients with acute pancreatitis. *Gastroenterol Res Pract*. 2021;2021:6699421.
19. Huang L, Chen C, Yang L, et al. Neutrophil-to-lymphocyte ratio can specifically predict the severity of hypertriglyceridemia-induced acute pancreatitis compared with white blood cell. *J Clin Lab Anal*. 2019;33:e22839.
20. Peri G, Introna M, Corradi D, et al. PTX3, A prototypical long pentraxin, is an early indicator of acute myocardial infarction in humans. *Circulation*. 2000;102:636–641.
21. Staubli SM, Oertli D, Nebiker CA. Laboratory markers predicting severity of acute pancreatitis. *Crit Rev Clin Lab Sci*. 2015;52:273–283.
22. Amálio SM, Macedo MA, Carvalho SM, et al. Mortality assessment in patients with severe acute pancreatitis: a comparative study of specific and general severity indices. *Rev Bras Ter Intensiva*. 2012;24:246–251.
23. Gurda-Duda A, Kuśnierz-Cabala B, Nowak W, et al. Assessment of the prognostic value of certain acute-phase proteins and procalcitonin in the prognosis of acute pancreatitis. *Pancreas*. 2008;37:449–453.
24. Park JM, Shin SP, Cho SK, et al. Triglyceride and glucose (TyG) index is an effective biomarker to identify severe acute pancreatitis. *Pancreatology*. 2020;20:1587–1591.
25. Tukiainen E, Kylänpää ML, Repo H, et al. Hemostatic gene polymorphisms in severe acute pancreatitis. *Pancreas*. 2009;38:e43–e46.
26. Cuthbertson CM, Christophi C. Disturbances of the microcirculation in acute pancreatitis. *Br J Surg*. 2006;93:518–530.
27. Radenkovic D, Bajec D, Ivancevic N, et al. D-dimer in acute pancreatitis: a new approach for an early assessment of organ failure. *Pancreas*. 2009;38:655–660.
28. Qiu Q, Nian YJ, Guo Y, et al. Development and validation of three machine-learning models for predicting multiple organ failure in moderately severe and severe acute pancreatitis. *BMC Gastroenterol*. 2019;19:118.
29. Balthazar EJ. Acute pancreatitis: assessment of severity with clinical and CT evaluation. *Radiology*. 2002;223:603–613.
30. Mortele KJ, Wiesner W, Intriere L, et al. A modified CT severity index for evaluating acute pancreatitis: improved correlation with patient outcome. *AJR Am J Roentgenol*. 2004;183:1261–1265.