



# OPEN Constructing a prediction model for acute pancreatitis severity based on liquid neural network

Jie Cao<sup>1,4</sup>, Shike Long<sup>2,4</sup>, Huan Liu<sup>1</sup>, Fu'an Chen<sup>1</sup>, Shiwei Liang<sup>1</sup>, Haicheng Fang<sup>1</sup> & Ying Liu<sup>1,3</sup>✉

Acute pancreatitis (AP) is a common disease, and severe acute pancreatitis (SAP) has a high morbidity and mortality rate. Early recognition of SAP is crucial for prognosis. This study aimed to develop a novel liquid neural network (LNN) model for predicting SAP. This study retrospectively analyzed the data of AP patients admitted to the Second Affiliated Hospital of Guilin Medical University between January 2020 and June 2024. Data imbalance was dealt with by data preprocessing and using the synthetic minority oversampling technique (SMOTE). A new feature selection method was designed to optimize model performance. Logistic regression (LR), decision tree (DCT), random forest (RF), Extreme Gradient Boosting (XGBoost), and LNN models were built. The model's performance was evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUC) and other statistical metrics. In addition, SHapley Additive exPlanations (SHAP) analysis was used to interpret the prediction results of the LNN model. The LNN model performed best in predicting AP severity, with an AUC value of 0.9659 and accuracy, precision, recall, F1 score, and specificity higher than 0.90. SHAP analysis revealed key predictors, such as calcium level, amylase activity, and percentage of basophils, which were strongly associated with AP severity. As an emerging machine learning tool, the LNN model has demonstrated excellent performance and potential in AP severity prediction. The results of this study support the idea that LNN models can be applied to early severity assessment of AP patients in a clinical setting, which can help optimize treatment plans and improve patient prognosis.

**Keywords** Liquid neural network, Severe acute pancreatitis, Predictive models, Machine learning

Acute pancreatitis (AP) is a common gastrointestinal disease with a worldwide incidence of approximately 4.9–73.4 per 100,000<sup>1</sup>. As the disease progresses, its severity and prognosis vary. In the world, about 20% to 30% of AP can develop into severe acute pancreatitis (SAP), which has a poor prognosis. Meanwhile, it has been shown that compared to mild AP (MAP), the mortality rate of SAP can be approximately ten times higher, accompanied by localized complications and organ failure, complicated treatment, and high disease burden<sup>2</sup>. Therefore, assessing the severity of AP in an early stage, adjusting the treatment plan, and intervening are conducive to improving the prognosis of AP<sup>3</sup>.

Many clinical scoring systems have been developed to predict disease severity, including the Ranson score, bedside index for severity in AP (BISAP), acute physiological and chronic health evaluation II (APACHE II), and CT severity index (CTSI). C-reactive protein (CRP) is also widely used as a single indicator to assess the severity of AP<sup>4</sup>. However, it is often difficult for a single indicator to fully reflect the complexity of the disease. As the first specialty scoring system, the Ranson score contains 11 indicators; despite being widely used, it takes 48 hours to complete and has a relative delay preventing an early and accurate assessment<sup>3,6</sup>. The BISAP score, which is favored for its simplicity and ease of accessibility, contains urea nitrogen, age, state of consciousness, pleural effusion, and systemic inflammatory response (SIRS). It has high sensitivity and positive predictive value<sup>7,8</sup>. However, the BISAP score does not adequately account for the effects of pancreatic inflammation on the gastrointestinal tract<sup>9</sup>. The APACHE II score, although widely used in the evaluation of patients with acute and critical illnesses, contains 15 complex parameters that are difficult to measure and focuses mainly

<sup>1</sup>Department of Gastroenterology, The Second Affiliated Hospital of Guilin Medical University, Guilin 541199, China.

<sup>2</sup>Guangxi University Key Laboratory of Unmanned Aircraft System Technology and Application, Guilin University of Aerospace Technology, Guilin 541004, China. <sup>3</sup>Guangxi Health Commission Key Laboratory of Glucose and Lipid Metabolism Disorders, The Second Affiliated Hospital of Guilin Medical University, Guilin 541199, China. <sup>4</sup>J. Cao and S. Long contributed equally to this work as co-first authors. ✉email: yingliu@glmc.edu.cn

on systemic conditions, with a lower degree of specificity than Ranson and BISAP<sup>10–12</sup>. The CTSI is primarily based on imaging studies, does not consider the SIRS, and is affected by the level of experience of the radiologist, losing its generalizability<sup>13</sup>. The limitations of these scoring systems suggest that no single scoring system can comprehensively assess the severity of AP.

As a branch of artificial intelligence, machine learning has been widely used in medicine due to its powerful computational and learning capabilities. Studies<sup>14–17</sup> have shown that machine learning has made remarkable achievements in constructing medical models, including disease diagnosis, clinical prognosis, and meta-analysis. Machine learning also plays a vital role in AP severity prediction. Pearce's team firstly utilized machine learning based on APACHE-II score and CRP metrics modeling to improve SAP prediction performance, and their model outperformed the traditional scoring system APACHE-II score in removing redundant features and identifying relevant features<sup>18</sup>. Xu Qiuping's team at Zhejiang University<sup>19</sup> established a low-latency standardized SAP scoring system using a deep confidence network. Sun et al.<sup>20</sup> developed the APSAVE model to analyze clinical variables through the random forest (RF) algorithm, which yielded an area under the receiver operating characteristic (ROC) curve (AUC) of 0.73, and combined with the Ranson score improved the model performance to 0.79. Jin et al.<sup>21</sup> compared the multilayer perception-artificial neural network (MPL-ANN) and partial least squares discrimination (PLS-DA) machine learning models in terms of their efficacy in predicting the severity of the disease of patients with AP and concluded that the MPL-ANN model has higher diagnostic and predictive capabilities. These studies suggest that machine learning techniques have significant AP severity prediction potential. However, as far as we know, no one has used a liquid neural network (LNN) model to predict disease severity in AP patients. This study aimed to develop the LNN model and compare it with traditional logistic regression (LR), decision tree (DCT), RF, and extreme gradient boosting (XGBoost) models to predict SAP. The innovation and significance of this study are:

1. LNN is applied to AP severity prediction for the first time. It can perform dynamic time series modeling and handle time series data effectively. It is more advantageous than traditional machine learning models in small sample and time series data processing.
2. A new feature selection method based on the AUC drive is proposed to optimize the data balance by combining it with the synthetic minority oversampling technique (SMOTE) technique. The most critical feature combinations for model prediction are screened out through nonparametric tests and correlation analysis. Compared with existing studies, this method improves model prediction accuracy, reduces redundancy and overfitting risk in the feature selection process, and enhances model efficiency and generalization ability.
3. The SHapley Additive exPlanations (SHAP) analysis revealed the importance of non-traditional biomarkers, such as calcium levels and basophil ratio, in predicting the severity of AP, which provides new potential targets for early clinical interventions and helps to optimize the therapeutic regimen. This study provides clinicians with a reliable tool to accurately assess the severity of AP patients at an early stage, improving patient prognosis and increasing survival.

## Materials and methods

### Patient characteristic information and selection criteria

It is a retrospective observational study that included all patients diagnosed with AP admitted to the Second Affiliated Hospital of Guilin Medical University between January 2020 and June 2024. It compiled data on the patients' routine examination indexes within 24 hours of admission. While collecting patient data, we formulated clear inclusion and exclusion criteria. The inclusion criteria were (1) meeting the diagnostic criteria for AP, (2) age  $\geq 18$  years, and (3) admission to the hospital with complete indicators of the required tests. The exclusion criteria were (1) age  $< 18$  years; (2) nursing or pregnant women; (3) patients with various malignant tumors; (4) patients with coagulation system disorders; (5) patients with chronic pancreatitis; and (6) patients with a large number of missing test results after admission. The definition criteria of SAP referred to the definition of SAP in the 2012 Atlanta classification standard<sup>4</sup>, as well as incorporating the actual situation of data collection. The patients' time to organ failure  $\geq 48$  hours during the disease was used as the predictive endpoint for SAP. Specifically: ① respiratory failure:  $\text{Pa O}_2/\text{Fi O}_2 \leq 300$ . ② circulatory failure: (systolic blood pressure  $< 90$  mm Hg or mean arterial pressure  $< 70$  mm Hg) and need to use vasoactive drugs. ③ Renal failure: blood creatinine  $\geq 170$   $\mu\text{mol/L}$  or  $> 3$  times of baseline or urine output  $< 0.5\text{ml/kg/h}$  for more than 24h or anuria for 12h. The human samples used in this study followed the principles of the Declaration of Helsinki, were approved by the Ethics Committee of the Second Affiliated Hospital of Guilin Medical University (NO.ZLXM-2024013), and informed consent was obtained from the patients and/or their legal guardians.

### Data preprocessing and feature selection

In addition to the basic information, the relevant data of the patients were mainly for several categories of tests, including electrolytes, amylase (AMY), liver function, coagulation function, renal function, cardiac enzymes, routine blood, lipids, and CRP, with a total of 105 characteristic data selected. Some features need to be vectorized in these feature data, mainly utilizing One-Hot coding. For example, in the patient's AP severity status feature, "0" means the patient's severe disease, and "1" means the patient's mild disease. In the gender feature, "1" represents male, and "2" represents female. The patient features are rationally represented in vector form by this processing step.

The feature variables with more than 40% missing data were deleted so that 105 types of feature data were deleted, leaving 64 features, as shown in Tables 1 and 2. The most commonly used method of filling in features with less than 40% missing data is the variance and mean method, which is suitable for missing variables with slight variance. However, the mean and variance interpolation method is no longer ideal for solving the problem of missing data in cases where individual differences in patients with AP are more pronounced. Therefore, the K

Variables	MAP (n = 585)	SAP (n = 137)	P value
Age (Years)	47 (36,62)	47 (37,66)	0.54
White Blood Cell Count (WBC) ( $10^9/L$ )	10.62 (7.39,14.24)	14.58±6.03	<b>0.000</b>
Red Blood Cell Count (RBC) ( $10^{12}/L$ )	4.53 (3.97,5.03)	4.78 (4.06,5.49)	<b>0.005</b>
Hemoglobin (HGB) (g/L)	135 (119,152)	142 (121,168)	<b>0.000</b>
Platelet Count (PLT) ( $10^9/L$ )	218 (163,269)	232 (168,282)	0.172
Neutrophils Percentage (NEU% ) (%)	79.4 (69.4,86.1)	86.3 (81,90.15)	<b>0.000</b>
Neutrophils Absolute Count (NEU#) ( $10^9/L$ )	8.31 (5.28,11.72)	12.59±5.75	<b>0.000</b>
Lymphocytes Percentage (LYM%) (%)	11.1 (6.3,17.95)	7.3 (4.3,10.75)	<b>0.000</b>
Lymphocytes Absolute Count (LYM#) ( $10^9/L$ )	1.23 (0.81,1.705)	1.01 (0.665,1.435)	<b>0.002</b>
Monocytes Percentage (MON%) (%)	6 (4.45,7.85)	5.1 (3.8,6.8)	<b>0.001</b>
Monocytes Absolute Count (MON#) ( $10^9/L$ )	0.64 (0.44,0.88)	0.8 (0.46,1.14)	<b>0.003</b>
Eosinophils Percentage (EOS%) (%)	0.4 (0.1,1.3)	0.1 (0.1,0.3)	<b>0.000</b>
Eosinophils Absolute Count (EOS#) ( $10^9/L$ )	0.05 (0.01,0.13)	0.01 (0.01,0.04)	<b>0.000</b>
Basophil Percentage (BAS%) (%)	0.2 (0.1,0.4)	0.2 (0.1,0.3)	<b>0.000</b>
Basophil Absolute Count (BAS#) ( $10^9/L$ )	0.03 (0.01,0.04)	0.02 (0.02,0.04)	0.688
Hematocrit (HCT) (%)	40.2 (45.5,44)	41.8 (36.75,48)	<b>0.002</b>
Mean Corpuscular Volume (MCV) (fL)	89.3 (84.95,92.7)	89.5 (85.55,93.7)	0.260
Mean Corpuscular Hemoglobin (MCH) (pg)	30.2 (28.6,31.7)	30.7 (29.2,32.4)	<b>0.010</b>
Mean Corpuscular Hemoglobin Concentration (MCHC) (g/L)	335 (326,345)	338 (326,352)	0.052
Red Cell Distribution Width - Coefficient of Variation (RDW-CV) (%)	13 (12,14)	13 (13,14)	<b>0.000</b>
Red Cell Distribution Width - Standard Deviation (RDW-SD) (fL)	41.4 (38.7,44.15)	42.8 (40.25,46.35)	<b>0.000</b>
Plateletcrit (PCT) (%)	0.23 (0.17,0.275)	0.24 (0.18,0.29)	<b>0.031</b>
Mean Platelet Volume (MPV) (fL)	10.2 (9.5,10.9)	10.3 (9.7,11)	0.174
Platelet Distribution Width (PDW) (%)	11.7 (10.1,13.15)	11.9 (10.4,13.45)	0.166
Large Platelet Ratio (P-LCR) (%)	26.8 (20.8,32.5)	27.2 (22.85,33.45)	0.181
CRP (mg/L)	6.28 (0.4,59.13)	68.49 (4.09,163.44)	<b>0.000</b>
Total Protein (TP) (g/L)	72.8 (67.75,76.6)	70.8 (64.85,76.2)	<b>0.034</b>
Albumin (ALB) (g/L)	41.5 (35.35,44.4)	36.3 (31.25,42.2)	<b>0.000</b>
Globulin (GLOB) (g/L)	31.1 (27.4,35.2)	33.5 (30.65,36.55)	<b>0.000</b>
Albumin/Globulin Ratio (A/G)	1.27 (1.04,1.475)	1.1 (0.91,1.32)	<b>0.000</b>
Total Bilirubin (TBIL) (umol/L)	11.7 (7.35,19.9)	15.5 (10.25,28.75)	<b>0.000</b>
Indirect Bilirubin (IBIL) (umol/L)	6.2 (4,9.7)	6.9 (3.9,10.65)	0.267

**Table 1.** Patient information and clinical characteristics (first 32 characteristics). In MAP group, 402 males and 183 females; In SAP group, 83 males and 54 females; No statistical difference in comparison of males and females between the two groups

Nearest Neighbor algorithm (KNN) is used in this study. This method is based on the assumption that “similar inputs have similar outputs.” By calculating the distance between a given data point and other points in the training dataset, the closest K neighbors are identified, and predictions are made based on the category or value of these neighbors.

Secondly, this study also normalizes the data by using the maximum and minimum feature variables to convert the raw data to a range between [0,1], thus eliminating the magnitude effect and making it comparable between different features. The maximum and minimum normalization method is

$$x' = \frac{x - \min}{\max - \min}$$

where  $x'$  is the feature value after normalization,  $x$  is the original feature value, min is the minimum of the feature, max is the maximum of the feature. When the maximum value is substituted into the above equation, 1 is obtained; when the minimum value is substituted into the above equation, 0 is obtained. Therefore, after feature normalization, the range of the eigenvalues of the patient data is between 0 and 1.

Category imbalance is a common problem in medical data. It refers to the fact that some categories have more data in a given dataset, and some have less. There is a massive gap between the categories with a larger number and the data categories with a smaller number. This problem causes the model to focus on various categories differently, affecting the final prediction effect and producing bias. Hence, the issue of data category imbalance needs to be dealt with before constructing the model.

To address the problem of category imbalance, this study uses SMOTE, which increases the sample size of the source data through data augmentation, thus allowing the model to achieve better generalization. SMOTE is a

Variables	MAP (n = 585)	SAP (n = 137)	P value
Direct Bilirubin (DBIL) (umol/L)	4.8 (2.6,9)	7 (4.15,14.95)	<b>0.000</b>
Alanine Aminotransferase (ALT) (U/L)	25.8 (13.1,62.95)	31.2 (13.2,84.55)	0.276
Aspartate Aminotransferase (AST) (U/L)	25.6 (17.85,52.05)	36 (21.25,93.85)	<b>0.001</b>
AST/ALT Ratio	1.02 (0.67,1.61)	1.32 (0.78,2.19)	<b>0.001</b>
Gamma-Glutamyl Transferase (GGT) (U/L)	66 (19,170.5)	62 (10.10,184.5)	0.473
Alkaline Phosphatase (ALP) (U/L)	77 (53,99.5)	70 (31.21,96)	0.063
Cholinesterase (CHE) (U/L)	8138 (4181,10586)	6460 (1534.66,9260)	<b>0.006</b>
Prealbumin (PA) (mg/L)	226.7 (95.15,303.2)	156.9 (45.59,276.2)	<b>0.004</b>
Adenosine Deaminase (ADA) (U/L)	8 (5,11.8)	8 (3.05,11)	0.169
Total Bile Acids (TBA) (umol/L)	1.5 (0.2,4.25)	0.8 (0.1,3.35)	<b>0.010</b>
Creatinine (Cr) (umol/L)	66 (51,81)	75 (57,111.5)	<b>0.000</b>
Urea (mmol/L)	4.2 (2.8,5.7)	5 (3.4,7.95)	<b>0.000</b>
Carbon Dioxide Combining Power ( $CO_2CP$ ) (mmol/L)	22.7 (19.2,24.8)	19.4 (16.2,22.9)	<b>0.000</b>
Creatine Kinase (CK) (U/L)	74 (35.5,133)	82 (38,150.5)	0.271
Creatine Kinase-MB (CK-MB) (ng/mL)	13 (7.95,19)	15.2 (8.65,22.25)	<b>0.031</b>
Lactate Dehydrogenase (LDH) (U/L)	213 (153,287)	317 (186,472.5)	<b>0.000</b>
Alpha-Hydroxybutyrate Dehydrogenase ( $\alpha$ -HBDH) (U/L)	154 (111.5,204)	208 (132.5,323)	<b>0.000</b>
AMY (U/L)	194 (77,564.25)	595 (139,1086.55)	<b>0.000</b>
Potassium (K) (mmol/L)	3.69 (3.41,3.935)	3.72 (3.56,4.01)	0.264
Chloride (Cl) (mmol/L)	101.5 (97.85,104.1)	100.7 (97.4,103.5)	0.220
Sodium (Na) (mmol/L)	137.7 (135.2,140)	136.3 (132.1,139)	<b>0.000</b>
Calcium (Ca) (mmol/L)	2.24 (2.12,2.32)	2.03 (1.71,2.18)	<b>0.000</b>
Triglycerides (TG) (mmol/L)	1.44 (0.37,7.515)	2.83 (0.67,12.81)	<b>0.010</b>
Cholesterol (CHOL) (mmol/L)	4.75 (3.65,6.315)	4.81 (3.49,8.03)	0.195
Low-Density Lipoprotein Cholesterol (LDL-C) (mmol/L)	1.99 (1.03,2.785)	1.23 (0.53,2.51)	<b>0.001</b>
High-Density Lipoprotein Cholesterol (HDL-C) (mmol/L)	0.85 (0.54,1.22)	0.6 (0.35,1.03)	<b>0.000</b>
Prothrombin Time (PT) (S)	13.1 (12.5,13.8)	13.7 (12.9,15.3)	<b>0.000</b>
International Normalized Ratio (INR)	1.03 (0.97,1.1)	1.09 (1.01,1.26)	<b>0.000</b>
Prothrombin Time Activity (PT%) (%)	94.55 (82.54,108.17)	85.25 (67.53,100)	<b>0.000</b>
Activated Partial Thromboplastin Time (APTT) (S)	34 (31.2,37.35)	35.5 (32.2,40.8)	<b>0.001</b>
Fibrinogen (FIB) (g/L)	3.65 (2.93,4.91)	4.6 (3.38,6.4)	<b>0.000</b>
Thrombin Time (TT) (S)	16.3 (15.3,17.1)	16.4 (15.35,17.7)	0.094

**Table 2.** Patient information and clinical characteristics (last 32 characteristics). In MAP group, 402 males and 183 females; In SAP group, 83 males and 54 females; No statistical difference in comparison of males and females between the two groups.

sophisticated oversampling method that generates synthetic instances from small population categories. Instead of replicating cases, it selects two or more similar instances (using a distance metric such as the Euclidean distance). It perturbs one example along the line segment connecting the cases.

To avoid SMOTE may introduce synthetic noise and lead to overfitting. This study incorporates the KNN algorithm to optimize the application of SMOTE. After repeated trials, we set the parameter  $k$  to 5, i.e., five nearest-neighbor samples of each minority class sample were selected to generate synthetic data points, thus balancing the data set. A mechanism for rejecting outliers was also developed. Specifically, any synthetic sample that deviates  $\pm 3\sigma$  (3 times the standard deviation) from the original data distribution is discarded. This approach minimizes noise while effectively balancing the data across categories.

This study mixed multiple feature selection methods to select appropriate features and improve the model prediction effect. Firstly, the AP patients were divided into SAP and MAP groups, and the features with differences between the two groups were screened by non-parametric test analysis for the next machine learning modeling. Among the selected differential features, correlation analysis was first used to derive the degree of correlation for each feature. The features were ranked in descending order according to the degree of relevance, and the AUC value in each case was calculated by increasing the features in that order. This allows direct visualization of the feature indicators that cause the AUC value to drop. Deleting the feature metrics that cause a decrease in AUC value can filter out the most representative feature metrics to improve the predictive relevance and reduce redundancy. In feature incremental selection, the AUC value under the ROC is used as a judgment indicator.

Compared with traditional techniques such as recursive feature elimination (RFE) and principal component analysis (PCA), the feature selection method in this study has obvious advantages, unlike RFE, which removes features based on model coefficients and may ignore some essential features. In contrast, the present feature selection method is driven by AUC, which is directly oriented to improve the model performance. In addition, compared with PCA, a linear dimensionality reduction technique, PCA may ignore the nonlinear relationship

between features. This feature selection method retains more information by selecting original features instead of transformed components. It reduces interference from redundant features, improving the model's ability to capture complex patterns in the data.

### LR model

LR is a probability-based linear classification model that linearly combines the features of the samples and then maps the result of the linear combination to a probability value between 0 and 1 via a sigmoid function. The advantage is that the model is simple and easy to implement and interpret, making it suitable for tackling linearly differentiable classification problems. However, LR has some disadvantages: it is prone to overfitting, especially when the variables are highly correlated. A linear relationship between variables is assumed, which limits its ability to handle complex data relationships. Moreover, the relationship between variables may not always be linear in the real world.

### DT model

DCT is a machine learning technique that constructs tree models by recursively selecting optimal features and segmentation points. It is used for classification and regression tasks. It starts at the root node and gradually splits down the tree until predefined stopping conditions are met, such as reaching the maximum tree depth, a threshold for the number of node samples, or a sufficiently high node purity. The algorithm evaluates all features and split points at each node and selects the features and values that best distinguish the data. The dataset is then split into subsets based on the features chosen and split points, and the process is repeated on each subgroup. DCT is intuitive, easy to interpret, and able to handle nonlinear relationships and high-dimensional data, but it is prone to overfitting and noise sensitivity and may ignore small patterns.

### RF model

RF uses techniques such as repeated sampling and feature random sampling to construct multiple DT and combine their predictions to improve model accuracy. Repeated sampling can build various models on a limited dataset to avoid overfitting problems. On the other hand, feature random sampling reduces the model's dependence on certain features and improves the model's generalization ability. The advantages are high accuracy and interpretability, handling high-dimensional data and nonlinear relationships, and a certain tolerance for abnormal data and missing values. However, its training time is longer, and the model complexity is higher, requiring larger computational resources when dealing with large-scale datasets.

### XGBoost model

XGBoost is a robust machine learning algorithm that belongs to ensemble learning in generalized supervised learning. It constructs robust models by integrating a series of predictions from simpler models (usually DCT) to reduce the bias-variance trade-off and enhance the overall model prediction power. The core of XGBoost is an iterative algorithm, where each iteration tries to minimize the overall prediction error by adding a new weak learner to correct the mistakes that appeared in the previous iterations. It has high accuracy and generalization ability and performs well when dealing with large-scale and high-dimensional data. However, it is computationally expensive and time-consuming, especially when applied to large datasets with many features. Compared with basic models such as LR or DCT, the interpretability of the XGBoost model is poor.

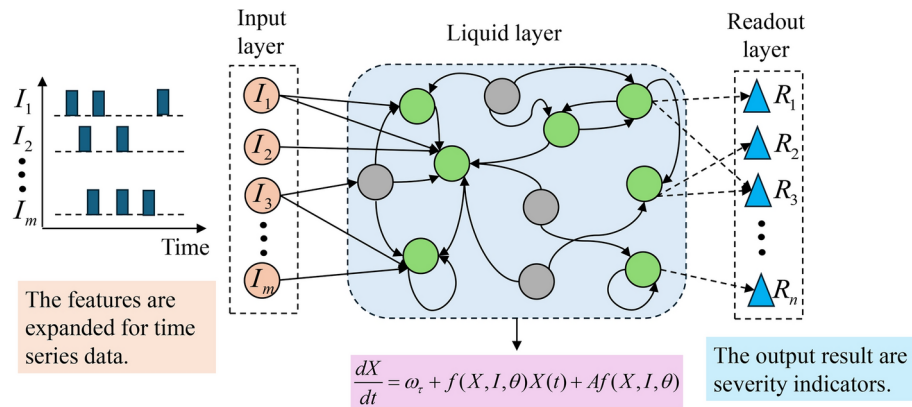
### LNN model

LNN is a sophisticated neural network inspired by the workings of the human brain, capable of processing data sequentially and adapting to changes in real time. LNN is a time-continuous Recurrent Neural Network. LNN not only processes the inputs sequentially but also retains the memory of the past inputs, adapts its behavior to new inputs, and can process variable-length inputs, significantly improving its understanding of the task. This adaptability of LNN gives it the ability to continuously learn and adapt to changes in the environment, especially in processing time series data, showing higher efficiency and stronger performance than traditional neural networks.

The main reason for choosing LNN is its unique advantage in handling dynamic time series and small sample data. Compared with traditional deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM), LNN is modeled by continuous-time differential equations, enabling it to capture nonlinear relationships in the data more flexibly. It also maintains a low parameter complexity, effectively reducing the risk of overfitting. This is especially important in small sample scenarios in the medical field. Its principle framework is shown in Fig. 1.

The structure contains three main layers: the input layer, the liquid layer, and the output layer. The input layer is the input to the network, which is provided with all the input data on which we wish to train the model. This layer feeds the input data to the liquid layer. The liquid layer contains a large recurrent network of neurons. They are initialized with synaptic weights, transforming the input data into a rich nonlinear space. The output layer consists of output neurons which receive information from the liquid layer. The most crucial is the liquid layer, which differs from traditional neural networks' hidden layer. It represents the continuous dynamics approximated by ordinary differential equation (ODE) and a more complex time step. Thus, the intermediate liquid layer states are obtained by solving the ODE. For example, in Fig. 1, the features are expanded into time-series data for a one-dimensional time-series input signal  $I(t)^{m \times 1}$  with  $m$  features for a given time step of  $t$ . A hidden state vector  $X^{(D \times 1)}(t)$  with  $D$  hidden units, and a time-constant parameter vector  $\omega_\tau^{(D)}$ . The liquid layer state of the LNN can be transformed into a solution solution for solving the following differential equation, which is:





**Fig. 1.** LNN schematic diagram.

$$\frac{dX}{dt} = \omega_\tau + f(X, I, \theta)X(t) + Af(X, I, \theta)$$

where  $\omega_\tau$  is a bias vector,  $f$  is a nonlinear liquid function with parameter  $\theta$ , and parameters  $A$  and  $\theta$  are system parameters, indicating that the coefficients of  $X(t)$  can vary as a function of state and input. The final severity classification  $R_i$  can be derived by connecting the hidden state vectors in an omnidirectional way. The expression is

$$R_i = \sigma[W \cdot X(t) + B]$$

where  $W$  is the weight matrix of the hidden state and the output result, and  $B$  is the bias matrix.  $\sigma[\cdot]$  is the activation function, usually a sigmoid function.

### Model interpretability

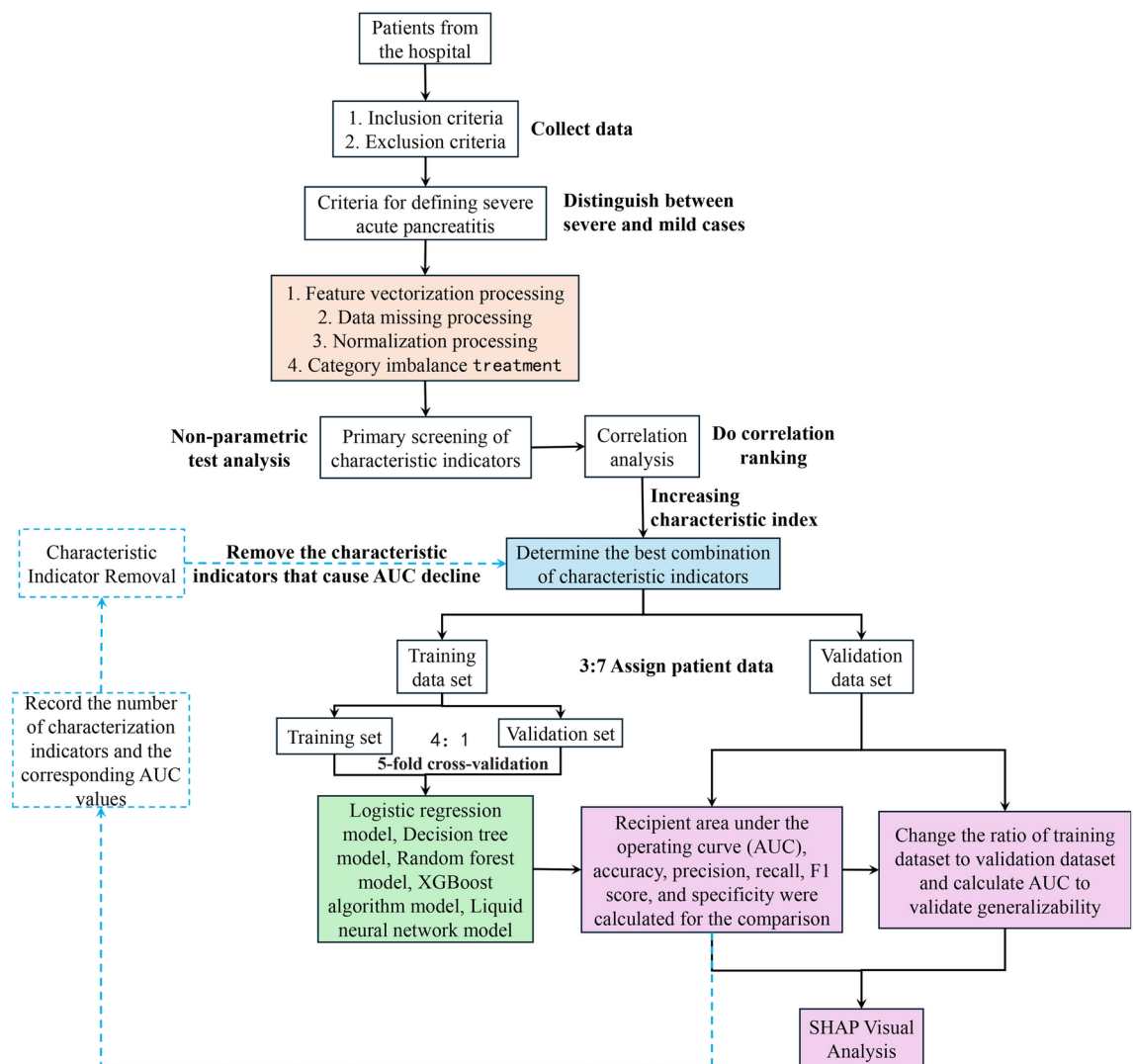
Besides predicting the severity of AP, another essential goal is using SHAP to identify the most critical factors and their contribution to the prediction. SHAP is a method for interpreting the prediction results based on the theory of Shapley values, which provides global and local interpretability of the model by decomposing the prediction results into each feature's influence. The core idea is to assign contributions to different features, calculate the Shapley value of each feature, and multiply it with the feature value to get the contribution of that feature to the prediction result. Graphical and quantitative interpretation results can be generated to help users explain the model's decision-making process.

### Overall prediction model construction process

The entire process of constructing a prediction model in this study is summarized in Fig. 2. We focus on the analysis of developing a model to predict AP severity based on LNN. Four algorithms, LR, DT, RF, and XGBoost, were chosen as comparison models. In this experiment, we filter the data according to the inclusion and exclusion criteria and then vectorize the data with features (mainly using One-hot coding for processing). The KNN algorithm is used for missing data, which is then transformed to data between 0 and 1 by the maximum and minimum normalization methods after the completion. In the data collected in this study, there were significantly fewer severely ill patients than sick, mild patients, so data imbalance existed. Here, the SMOTE technique was used for data resampling to balance each other. Based on the completion of data preprocessing, we randomly divided the dataset into a 70% training set and a 30% test set to prevent data information leakage. The training set is used to construct the model and adjust the hyperparameters, and the test set is used to evaluate the model's generalization performance.

Not all available features are used to train the model for the feature selection of each model. The best subset of features needs to be selected for each model. In this study, a novel feature selection method was designed. Firstly, AP patients were divided into SAP and MAP groups, and feature indicators with differences between the two groups (e.g., Table 1 and Table 2) were screened by nonparametric test analysis for the following machine learning modeling step. Among the selected differential feature indicators, correlation analysis was first used to derive the correlation degree of each feature indicator, as shown in Fig. 3. Based on the correlation degree, the features are ranked in decreasing order, and different combinations of features are constructed by incrementing the features step by step according to this order. The model under each combination is trained through the training set, and the AUC value is obtained through the validation set. This allows the feature metrics to be directly visualized, which causes the AUC value to drop. The most representative feature metrics can be filtered out by deleting the feature index that causes the AUC value to decrease. Finally, the best feature combinations of the five models are shown in Table 3, where the best features are 28 for the LR model, 22 for the DCT model, 29 for the RF model, 34 for the XGBoost model, and 27 for the LNN model.

All five models were implemented using Python, and the parameters were left at their default values for the LR model. For the DCT model, the maximum depth of the tree was set to 10, the minimum number of samples



**Fig. 2.** The flow chart of this study.

required to split the internal nodes was set to 2, the minimum number of samples needed to separate the leaf nodes was set to 4, and the segmentation quality measure was set to 'Gini'. For the RF model, the number of trees is set to 50, the maximum depth of the tree is set to 2, the minimum number of samples needed to split internal nodes is set to 2, and the minimum number of samples required for leaf nodes is set to 4. For the XGBoost model, the number of trees is set to 30, the maximum depth of the tree is set to 1, the learning rate is set to 0.1, the proportion of samples used for training and the proportion of features used to construct the tree are both set to 1, and the evaluation metric is set to "gloss." For the LNN model, the size of the intermediate liquid layer is set to 100, and the size of the input layer is the same as the number of selected feature combinations, except that a time series is added. The output layer is set to 1 because this study is a binary classification problem. The other parameters are set to initial values, the learning rate is set to 0.01, and the maximum number of iterations is set to 300.

Models are trained using 5-fold cross-validation to avoid model overfitting. AUC value, accuracy, precision, recall, F1 score, and specificity were calculated based on the test set to evaluate the five models' performance. Each model performance metric is calculated after SMOTE resampling and selecting the best combination of features. Secondly, model generalization performance evaluation was performed by gradually changing the proportion of data in the training set. Finally, a SHAP analysis of the LNN model was conducted to assess which inputs were most important for generating predictions based on the LNN model. The SHAP plots ranked the features predicted by the model in decreasing order of importance from top to bottom.

### Statistical analysis

Statistical analysis of patient data following preprocessing was conducted using SPSS version 25.0 software. Prior to analysis, the distribution of continuous variables was assessed for normality. Normally distributed data were presented as mean  $\pm$  standard deviation, while skewed data were reported as median (interquartile range). Clinical and biochemical characteristics were analyzed utilizing chi-square tests or nonparametric tests (Mann-

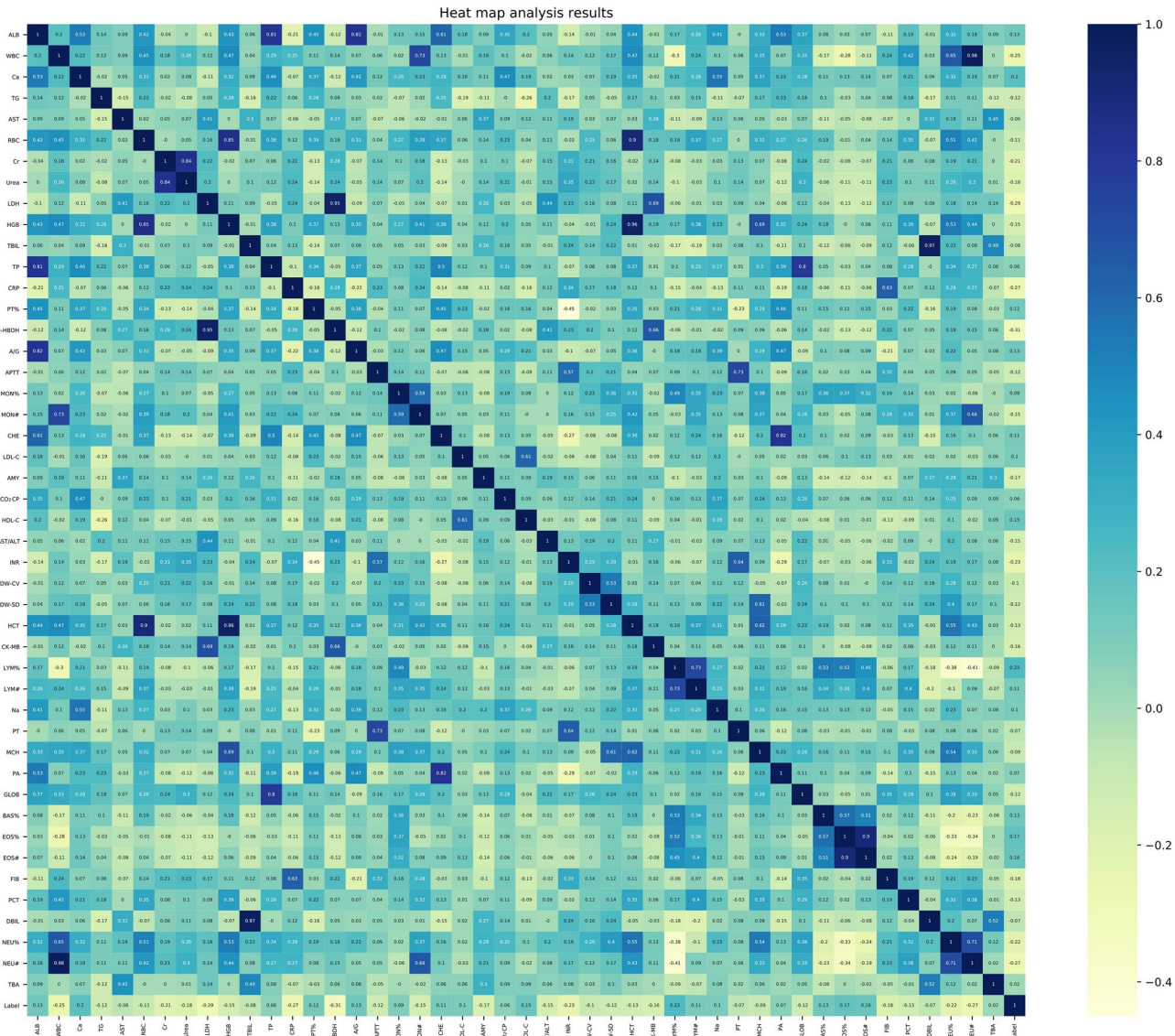


Fig. 3. Correlation analysis heatmap.

Models	Best feature combination
LR	“α-HBDH”, “LDH”, “CRP”, “NEU#”, “INR”, “LYM%”, “NEU%”, “Ca”, “Urea”, “AMY”, “EOS%”, “EOS#”, “HGB”, “ALB”, “BAS%”, “PCT”, “TG”, “APTT”, “RDW-SD”, “GLOB”, “CHE”, “LYM#”, “Na”, “MON%”, “MCH”, “TBIL”, “TP”, “CO <sub>2</sub> CP”
DCT	“α-HBDH”, “LDH”, “NEU#”, “WBC”, “Ca”, “FIB”, “AMY”, “CK-MB”, “MON#”, “HDL-C”, “A/G”, “HCT”, “BAS%”, “TG”, “RDW-SD”, “CHE”, “LDL-C”, “MCH”, “DBIL”, “AST”, “TP”, “CO <sub>2</sub> CP”
RF	“α-HBDH”, “LDH”, “CRP”, “NEU#”, “WBC”, “INR”, “NEU%”, “Cr”, “Ca”, “AMY”, “EOS%”, “CK-MB”, “HGB”, “HDL-C”, “A/G”, “BAS%”, “TG”, “APTT”, “RDW-SD”, “GLOB”, “CHE”, “RDW-CV”, “MON%”, “TBIL”, “PA”, “DBIL”, “TP”, “CO <sub>2</sub> CP”, “TBA”
XGBoost	“α-HBDH”, “LDH”, “CRP”, “NEU#”, “WBC”, “INR”, “NEU%”, “Cr”, “Ca”, “Urea”, “AMY”, “EOS%”, “CK-MB”, “EOS#”, “HGB”, “HDL-C”, “ALB”, “A/G”, “BAS%”, “PCT”, “PT%”, “APTT”, “RDW-SD”, “GLOB”, “CHE”, “LYM#”, “RDW-CV”, “MCH”, “TBIL”, “PTT”, “DBIL”, “AST”, “TP”, “CO <sub>2</sub> CP”
LNN	“α-HBDH”, “CRP”, “NEU#”, “WBC”, “INR”, “Cr”, “Ca”, “Urea”, “FIB”, “AMY”, “EOS%”, “HGB”, “HDL-C”, “A/G”, “BAS%”, “TG”, “APTT”, “RDW-SD”, “RBC”, “CHE”, “LDL-C”, “RDW-CV”, “MON%”, “MCH”, “PT”, “DBIL”, “CO <sub>2</sub> CP”

Table 3. The optimal combination of features for the five models.

Whitney U), with initial screening for differences set at a significance level of  $P < 0.05$ . The entire modeling process was implemented in Python version 3.8.5, employing the scikit-learn package version 0.24.1 for RL, DCT, and RF analyses; XGBoost model construction utilized the xgboost package version 1.4.2; and LNN model development, along with CNN model and LSTM model construction, was carried out using the torch package version 1.10.0. Additionally, SHAP analysis was performed using the shap package version 0.37.0. All data and



source code pertaining to this study are openly accessible on GitHub at <https://github.com/longshike/LNN-for-SAP-Prediction>.

## Results

### Univariate analysis of patient data

We collected data from a total of 732 AP patients according to the inclusion and exclusion criteria, of which 137 had SAP as defined by our Atlanta criteria and the remaining 585 had MAP. The data contained a total of 105 types of feature information, and 64 types of feature information remained after excluding features with a high number of missing values. These 64 types of feature information were grouped according to SAP and MAP, and then analyzed using chi-square test or non-parametric test (Mann-Whitney U), with  $P < 0.05$  as the criterion for whether there was statistical variability or not, which in turn led to the initial screening of the data. Table 1 and Table 2 summarize the results of the 64 characteristics analyzed. Among them, in the blood routine indicators, WBC, RBC, HGB, NEU%, NEU#, LYM%, LYM#, MON%, MON#, EOS%, EOS#, BAS%, HCT, MCH, RDW-CV, RDW-SD, and PCT were statistically different. In liver function indices, TP, ALB, GLOB, A/G, TBIL, DBIL, AST, AST/ALTRatio, CHE, PA and TBA were statistically different. In renal function indices, Cr, Urea and  $CO_2CP$  were statistically different. In cardiac enzymes, CK-MB, LDH and  $\alpha$ -HBDH were statistically different. In electrolyte indicators, Na and Ca were statistically different. Among lipid indicators, TG, LDL-C and HDL-C were statistically different. Among coagulation indicators, PT, INR, PT%, APTT and FIB were statistically different. There was also a statistical difference in CRP and AMY.

### Feature selection for different models

The 46 characteristics indicators that differed were analyzed univariately to enter the model construction. Examining the correlation of these 46 features produces a heat map of the correlation magnitude between the features and the final severity, as shown in Fig. 3. These features are sorted according to the correlation from the largest to the smallest, and a subset of 46 feature combinations can be constructed by increasing the number in order. Then, five models are built based on different feature combinations, and different AUC values are calculated. In this way, the relationship between the AUC value and the number of feature combinations under the five models can be obtained, as shown in Fig. 4. It can be seen that certain features can increase the AUC value rapidly, and some features can decrease it. The best feature combinations under five models can be obtained by eliminating the features that cause the AUC value to decrease under different models, as shown in Table 3. The best feature combinations of other models are XGBoost>RF>LR>LNN>DCT. Among them,  $\alpha$ -HBDH, NEU#, Ca, AMY, BAS%,  $CO_2CP$  are common to the optimal feature combinations of the five models, which also indicates the importance of these features for predicting the severity of pancreatitis.

### Comparative analysis of feature selection and oversampling on the performance of different models

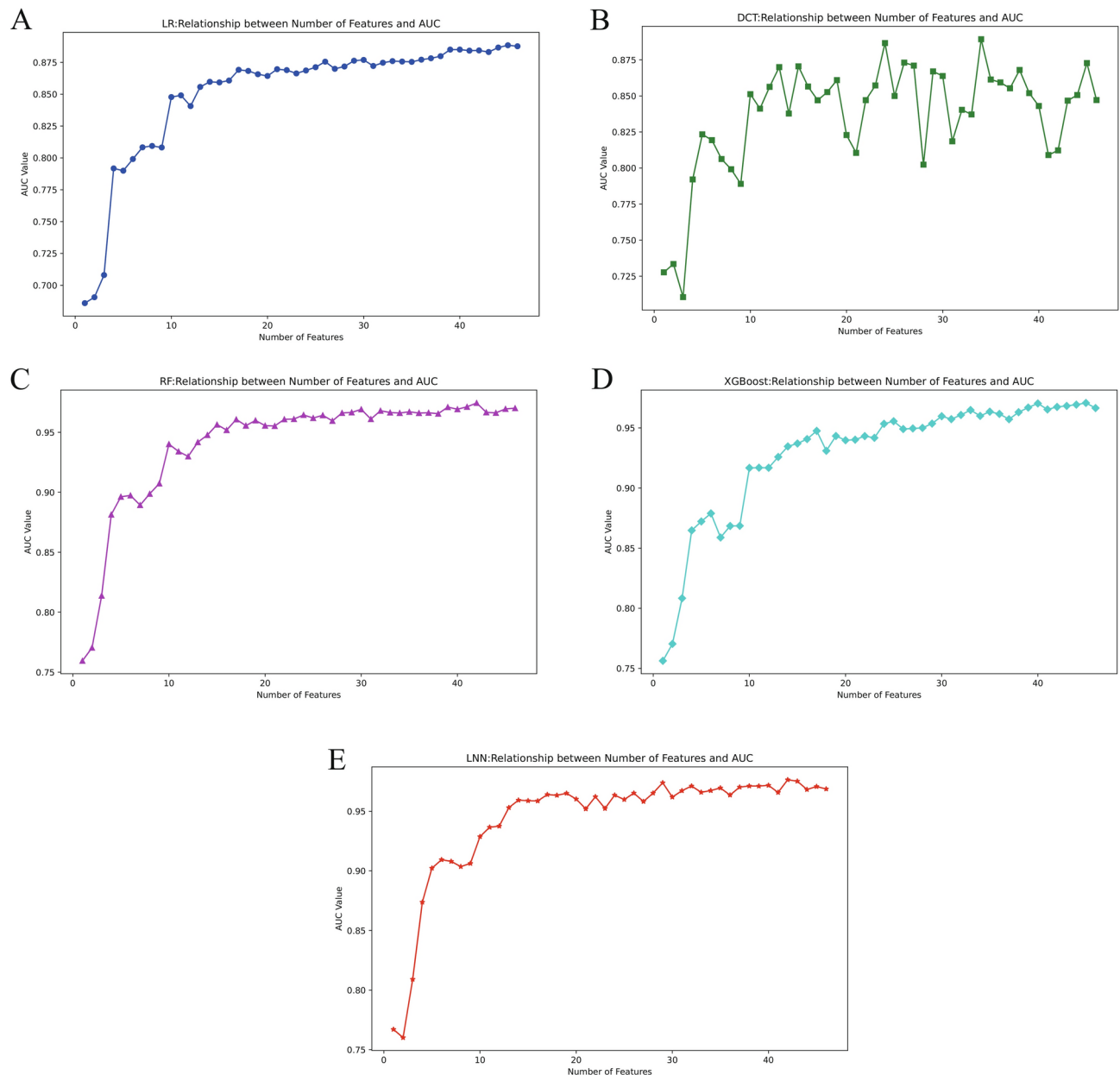
Figure 5 compares the five models (LR, DCT, RF, XGBoost, and LNN) regarding AUC values for the test data before and after feature selection. It can be seen that all models have improved performance after feature selection concerning the AUC values. For example, the AUC of the DCT model improves from 0.8318 to 0.8684. To further clarify the contribution of feature selection, Fig. 4 exhibits the quantitative relationship between the number of features and the AUC value. The results all show that selecting the optimal features combination can significantly enhance the model performance, e.g., the AUC value of the LNN model increases up to 0.27%. Once again, the effectiveness of the new feature selection method proposed in this study is verified in optimizing the model performance.

Feature selection is a process aimed at improving the performance of a model, and numerous variables such as data distribution, noise, and model complexity affect the ranking of feature importance. In some cases, there are interaction effects between features, and removing a feature may cause these effects to disappear. If too many features are removed, this may result in a loss of information and reduce the predictive potential of the model. In addition, feature selection reduces overfitting and improves model generalization by eliminating confusing or redundant features. Therefore, optimal feature selection for each model is an essential and effective process.

Figure 6 compares the AUC values of the five models on the test data before and after oversampling with SMOTE. After oversampling, the AUC of all models is improved, with the LNN showing the most significant increase, followed by the DCT. This demonstrates the efficacy of oversampling in managing unbalanced data. Notably, LNN shows the highest performance after oversampling among the five models. In classification problems, an unbalanced class distribution can significantly impact model performance. When there is a severe imbalance between the classes, the model will tend to predict the majority of the classes and thus ignore the minority of the categories. Oversampling techniques are utilized to increase the number of samples in the minority category, thereby enhancing the class balance and allowing the model to consider both the majority and minority categories equally. In addition, oversampling helps reduce overfitting by providing a larger and more diverse training dataset for the model to learn from. The reduction in overfitting further improves the performance of the model. In conclusion, by employing the oversampling technique and feature selection, the observations show that the class imbalance problem has been successfully mitigated, and the classification performance has been enhanced.

### Comparison of the five models' predictive performance

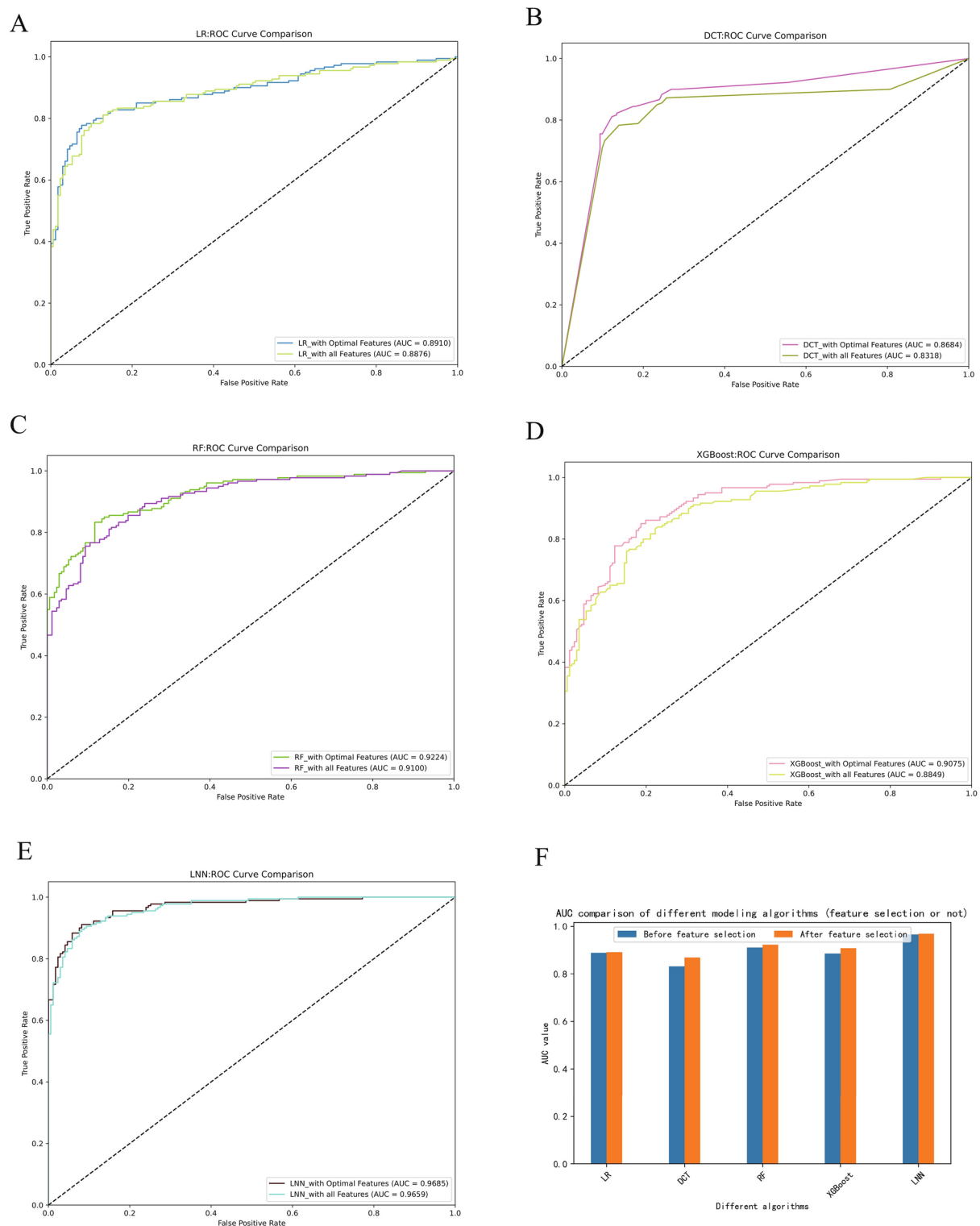
Figure 7 shows the change graphs of loss and accuracy in training the LNN model using the training set. The overall trend of the total loss error decreases, and the overall accuracy trend increases without abrupt changes. This shows that constructing the LNN model is correct. Figure 8 displays the ROC of the five models on the test set. The LNN has the highest AUC of 0.9659, followed by RF with an AUC of 0.9224, XGBoost with an AUC



**Fig. 4.** Plot of AUC values versus number of features under the five models. (A) LR model; (B) DCT model; (C) RF model; (D) XGBoost model; (E) LNN model.

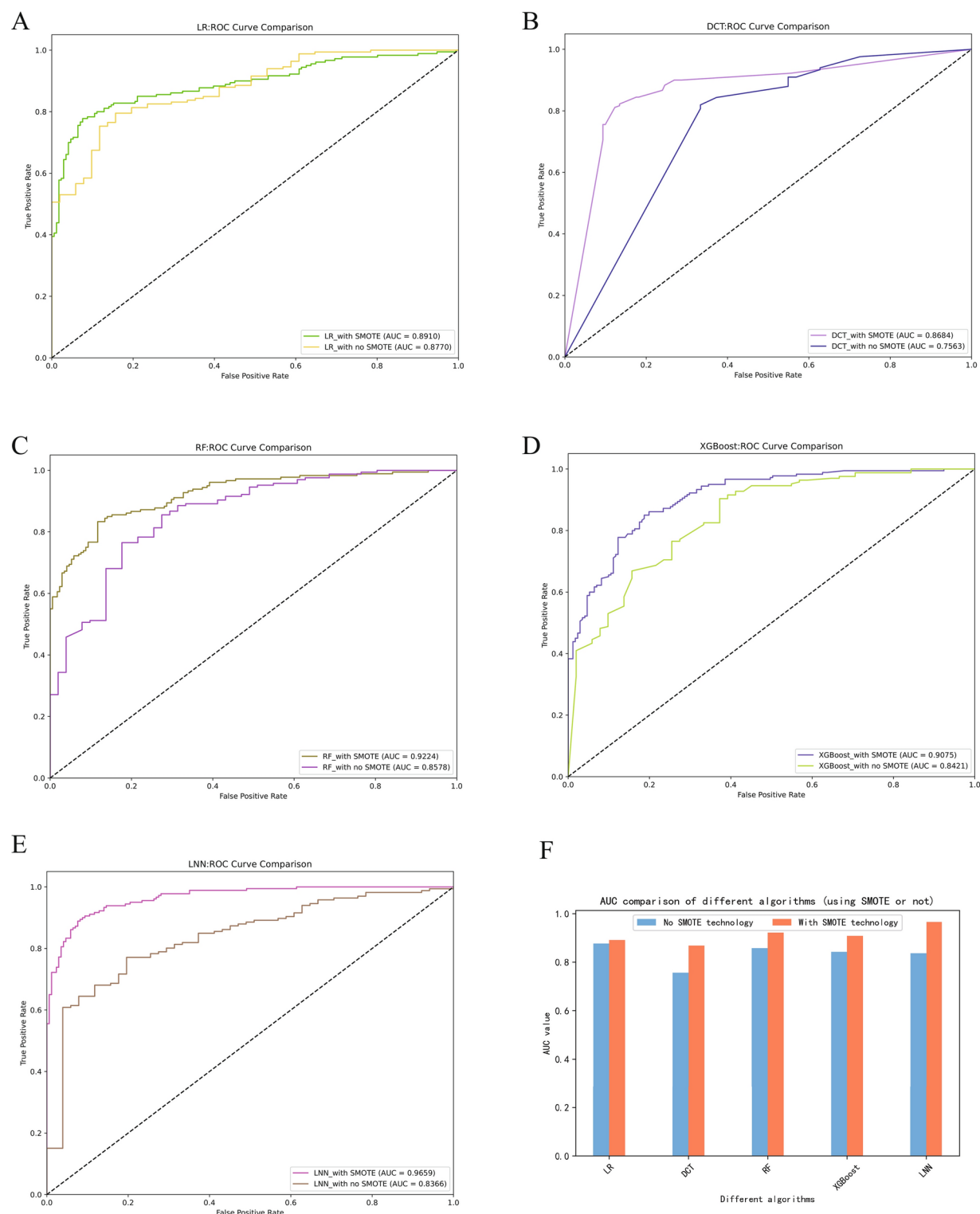
of 0.9075, LR with an AUC of 0.8910, and DCT with an AUC of 0.8684. Table 4 summarizes the performance of the models. Among the five classifiers, LNN has the best prediction performance with the highest accuracy of 0.9031, the highest precision of 0.9101, the highest recall of 0.9000, the highest F1 score of 0.9050, and the highest specificity of 0.9064, which indicates that LNN outperforms the other machine learning models in predicting AP severity.

To verify that the LNN model in this study provides better prediction for small sample data, numerical experiments are designed to gradually increase the training set percentage from 5% to 70% and validate it on the test set with AUC as the evaluation metric. AUC is applicable to measure the model's prediction performance with imbalanced classes. Numerical experiments are conducted to compare with LR, DCT, RF, and XGBoost models with different training set proportions, and the experimental results are shown in Table 5. As shown in Table 5, the LNN model proposed in this paper is superior to the other models, especially when the training set sample size is small. For example, when only 5% of the data are trained, the AUC of the method proposed in this paper reaches 0.8447. At the same time, the other models are still slightly inferior to the prediction effect due to fewer training sample instances. Meanwhile, with the increase of the training set proportion, the prediction effect AUC increases from 0.8447 to 0.9659 because the model has more sample data for training, and it can learn more information and predict better.



**Fig. 5.** Comparison of AUC values before and after feature selection under five models. (A) LR model; (B) DCT model; (C) RF model; (D) XGBoost model; (E) LNN model; (F) Comparison bar chart.

To better validate the advantages of LNN over traditional deep learning models, a comparison experiment of LNN with LSTM and CNN is designed. Figure 9A demonstrates that in the small sample case (only 5% of the training set), with the same feature selection data, LNN achieves an AUC value of 0.8447, which is significantly superior to LSTM's 0.7421 and CNN's 0.7503. even when 70% of the training set is used, LNN maintains the lead with 0.9659, which is 0.1767 and 0.0595 higher than LSTM and CNN, respectively, as shown in Fig. 9B. The



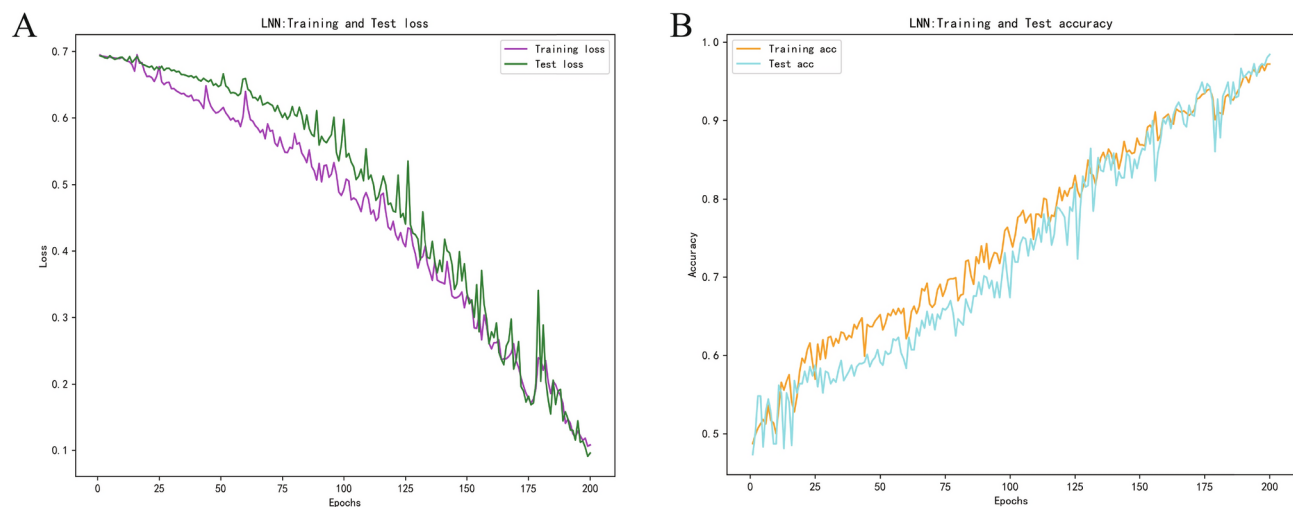
**Fig. 6.** Comparison of AUC values before and after SMOTE oversampling under five models. (A) LR model; (B) DCT model; (C) RF model; (D) XGBoost model; (E) LNN model; (F) Comparison bar chart.

results show that LNN has significant advantages in processing clinical dynamic data and real-time prediction, which is especially suitable for small-sample and time-series data analysis scenarios in the medical field.

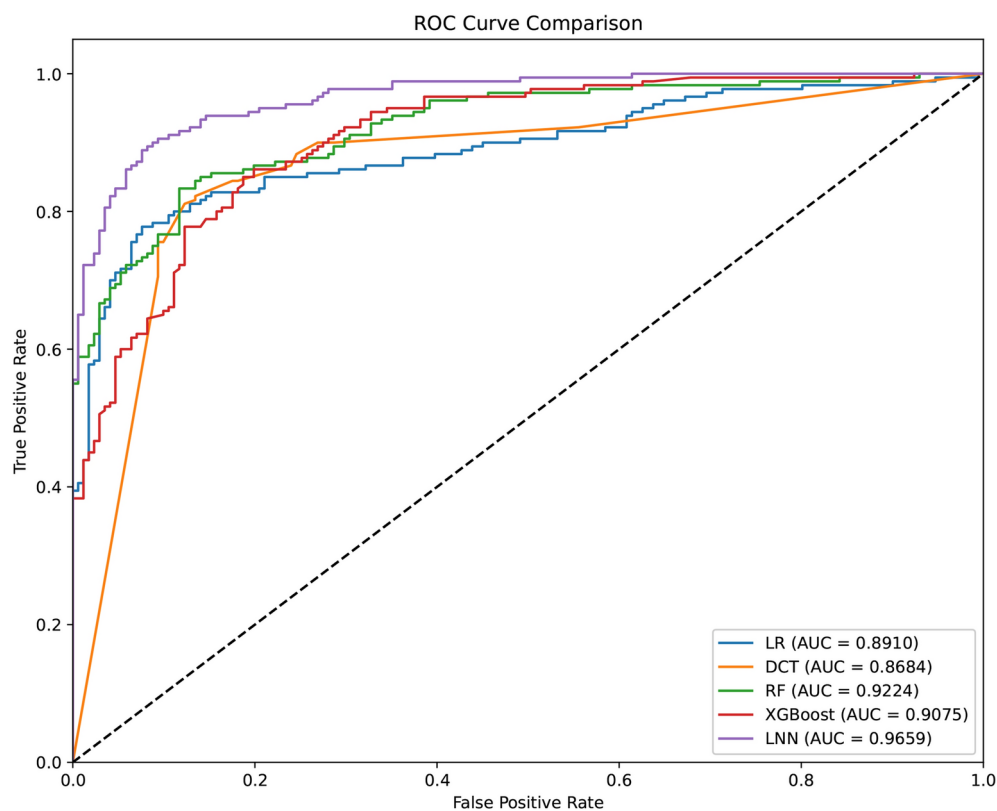
### Explaining LNN model

Figure 10 illustrates the SHAP contribution map based on the LNN model, with the 27 feature variables for which the model was constructed visually ordered according to their importance. The closer the value of a





**Fig. 7.** Loss and accuracy variation plots for training LNN models. (A) Loss variation chart; (B) Accuracy change chart.



**Fig. 8.** Comparative ROC charts for the five models.

variable is to 1, the more likely a patient is to develop SAP. It can be seen that the top 10 essential feature variables for the LNN model are sorted as Ca, AMY, BAS%,  $CO_2CP$ , EOS%,  $\alpha$ -HBDH, A/G, HDL-C, TG, and CRP. The most essential features are standard to the best feature combinations in Table 3. This also indicates the rationality and correctness of feature selection.

## Discussion

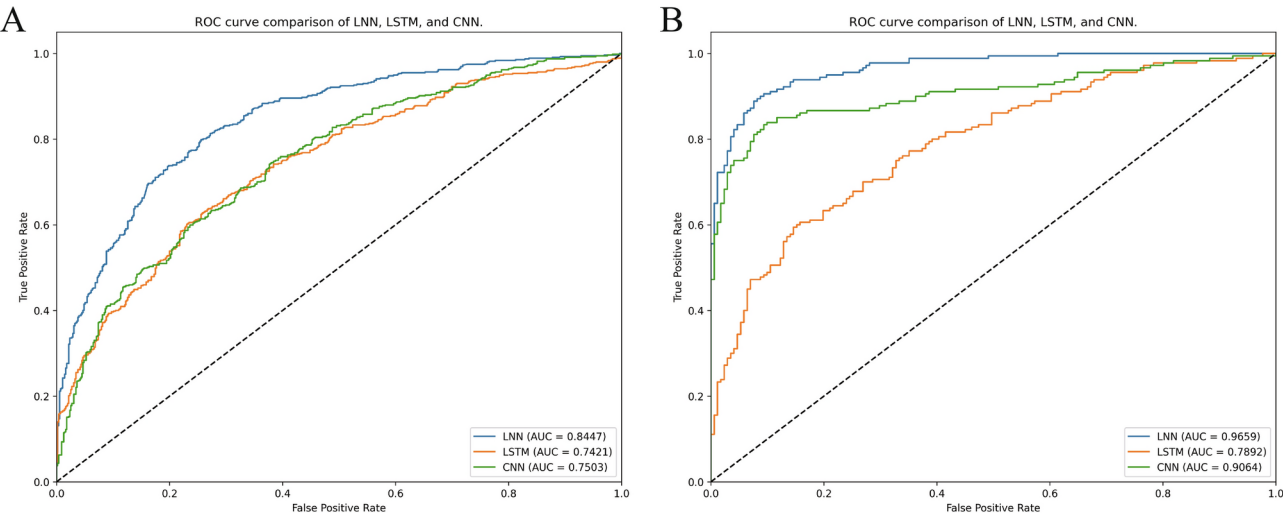
In this study, we developed an LNN-based predictive model to determine whether patients with AP developed SAP and compared LR, DCT, RF, and XGBoost models. In addition, this study developed a method for data preprocessing and feature selection to address the problem of imbalanced clinical data and small samples.

Models	Accuracy	Precision	Recall	F1 score	Specificity	AUC value
LR	0.8405	0.8780	0.8000	0.8372	0.8830	0.8910
DCT	0.8433	0.8655	0.8222	0.8433	0.8655	0.8684
RF	0.8433	0.8788	0.8056	0.8406	0.8830	0.9225
XGBoost	0.8034	0.8033	0.8167	0.8099	0.7895	0.9075
LNN	<b>0.9031</b>	<b>0.9101</b>	<b>0.9000</b>	<b>0.9050</b>	<b>0.9064</b>	<b>0.9659</b>

**Table 4.** Overall performance metrics for the five models.

Models	Training set							
	5%	10%	20%	30%	40%	50%	60%	70%
LR	0.8238	0.8345	0.8627	0.8613	0.8644	0.8744	0.8923	0.8910
DCT	0.7508	0.7693	0.8030	0.8295	0.8220	0.8477	0.8321	0.8684
RF	0.8608	0.8770	0.8997	0.9005	0.9092	0.9189	0.9236	0.9225
XGBoost	0.8106	0.8507	0.8609	0.8641	0.8714	0.8800	0.8840	0.9075
LNN	<b>0.8447</b>	<b>0.9002</b>	<b>0.9142</b>	<b>0.9413</b>	<b>0.9530</b>	<b>0.9564</b>	<b>0.9730</b>	<b>0.9659</b>

**Table 5.** AUC values for different models with different scaled training sets.

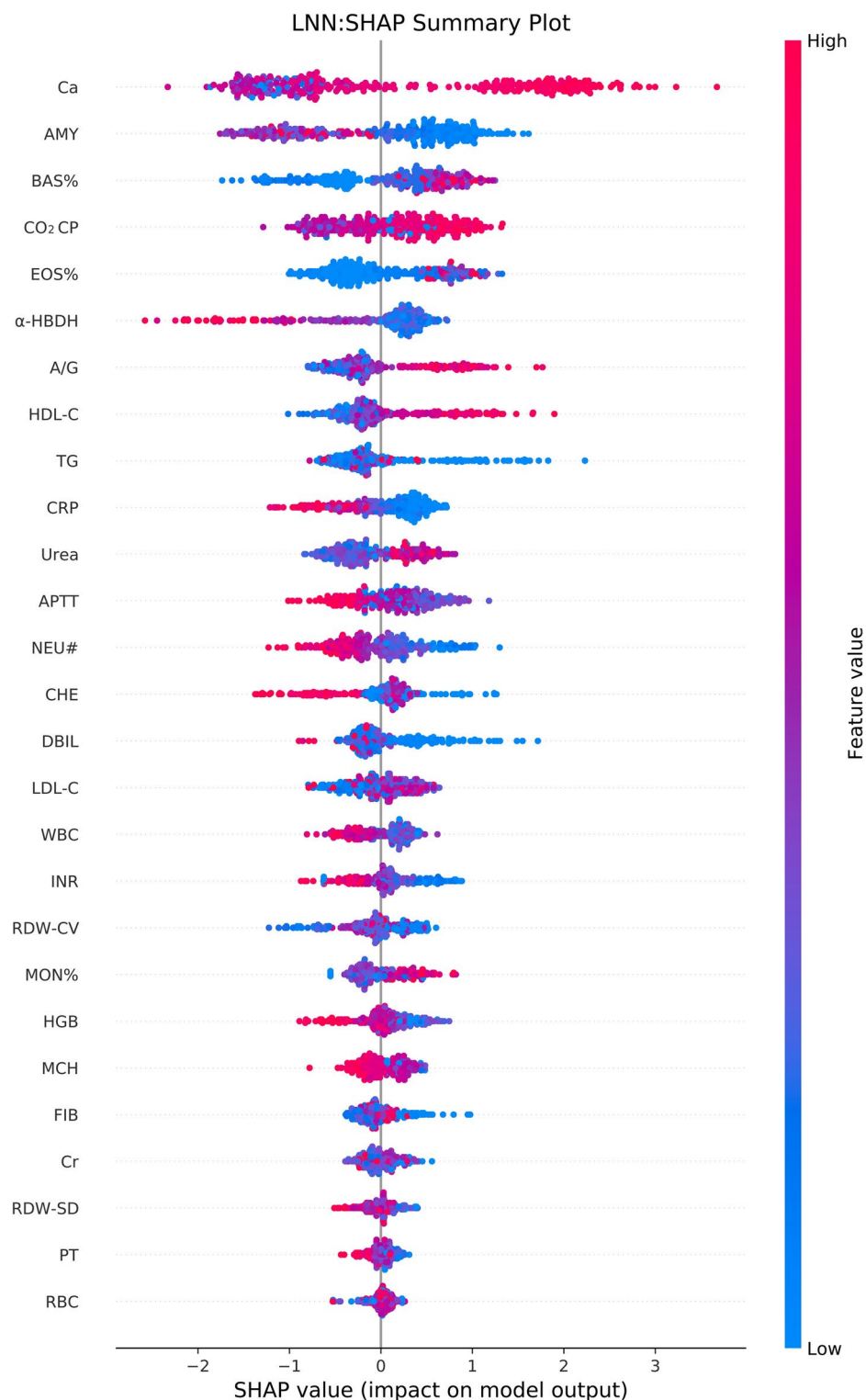


**Fig. 9.** Comparison of AUC values for three models, LNN, CNN, and LSTM, with different training set proportions. (A) 5% training set; (B) 70% training set.

A SHAP visualization of the LNN model’s features was subsequently conducted to derive the degree of each feature’s influence. The LNN-based model showed strong performance with an AUC above 0.90 and precision, accuracy, recall, F1 parameter, and specificity above 0.90. These results suggest that using LNN in a clinical setting can improve early severity prediction in AP patients.

Accurately predicting the severity of AP is critical in determining whether a patient needs to be admitted to the hospital, especially for patients who may develop SAP. Patients with SAP require intensive care and prompt medical interventions, including rapid fluid resuscitation, enteral feedings, and pain management, which need to be supported by hospital resources<sup>22,23</sup>. Additionally, in patients with necrotizing SAP, surgical intervention may be required<sup>23</sup>. So, early identification of patients with SAP is crucial to ensure that they can receive the necessary treatment and care on time, which can help improve treatment outcomes and patient survival<sup>8</sup>.

Given the importance of classifying patients with AP for early risk, several scoring systems have been developed and applied. In exploring scoring systems for assessing AP severity, we found that each system has unique limitations. Despite its long history, the Ranson score requires a long time to collect data, resulting in an inability to assess disease severity at an early stage<sup>5,6</sup>. The BISAP score, although simple and easy to use, is limited in terms of prognostication because it fails to adequately consider the impact of pancreatic inflammation on the gastrointestinal tract<sup>9</sup>. The APACHE II score, although comprehensive, is less specific in AP specialty assessment due to its complexity and focus on systemic conditions<sup>10–12</sup>. The limitations of the CTSI are its neglect of the systemic inflammatory response, the need for large-scale equipment in primary care hospitals, and reliance on



**Fig. 10.** LNN-based SHAP analysis graph.

the experience of radiologists, which limits its widespread clinical use<sup>13</sup>. These limitations suggest that although existing scoring systems are helpful to some extent in assessing the severity of AP, none of them can fully cover the multifaceted features of the disease. Therefore, developing new scoring systems to determine the AP severity more comprehensively and at an earlier stage is essential for improving the prognosis of patients.

As machine learning plays a significant role in the medical field, an increasing number of algorithms are being developed. Medical researchers are no longer limited to individual algorithms but are comparing multiple algorithms to find optimal solutions. New algorithms are also explored to augment the samples, thus solving

the real-world problems of small sample sizes, poor data quality, and unbalanced classification. Thapa's team<sup>24</sup> trained models using LR, XGBoost, and neural network algorithms, which yielded AUC of 0.780, 0.921, and 0.811 for the three models, which is a substantial improvement compared to the traditional scoring system. Kui et al.<sup>25</sup> designed a simple application based on the XGBoost algorithm to identify patients at high risk of SAP with a prediction accuracy of 89.1% and an average AUC of 0.81. Hong et al.<sup>26</sup> from Wenzhou Medical College developed an interpretable RF model to predict SAP. Yin et al.<sup>27</sup> from the First Affiliated Hospital of Suzhou University predicted AP severity at an early stage by using an automated machine learning (AutoML) technique, and the results showed that the model developed using XGBoost achieved the highest specificity value of 0.98 in the test set and the highest accuracy of 0.958. Yuan's team<sup>28</sup> used five machine learning algorithms by analyzing clinical data from 5460 AP patients, and the results suggested that the model based on the XGBoost algorithm achieved the highest AUC in the external test set. Zhou et al.<sup>29</sup>, Department of Gastroenterology, Affiliated Hospital of Yangzhou University, compared and developed five machine learning algorithms for constructing AP severity prediction models, and the results showed that XGBoost performed better in predicting SAP. Luo et al.<sup>30</sup> used five machine learning algorithms to construct a prediction model for SAP and compared it with a classical scoring system, and the results showed that the model based on the RF algorithm performed the best, with an AUC of 0.961. Ma et al.<sup>31</sup> compared the effectiveness of four machine learning techniques in predicting AP severity, and the results indicated that the XGBoost model performed the AP best with an AUC of 0.906. These findings further confirm the potential and value of machine learning in improving the accuracy of AP severity prediction.

Although RF and XGBoost algorithms perform better in predicting AP severity, these traditional machine learning methods are supervised learning and have some inherent limitations. The accuracy of these models is highly dependent on large-scale and high-quality datasets, and the prediction accuracy of the models suffers in the presence of insufficient sample sizes or imbalanced data categories. Besides, improper tuning of model parameters may also lead to overfitting, which reduces the model's generalization ability. Although deep learning has been widely used in medical image analysis due to its automatic feature extraction capability and ability to process complex data, it is still less used in AP severity prediction. Deep learning also faces the same dependence on large datasets, and the setting of hyperparameters becomes more complicated as the depth of the model increases. Complex neural network training tends to lead to overfitting on small sample datasets, making the model memorize the training samples instead of learning the underlying patterns. Therefore, to increase the accuracy of AP severity prediction models, future research needs to comprehensively consider the issues of sample size, feature selection, and data imbalance. It also balances the complexity, implementability, dynamic learning ability, generalization ability, robustness, and interpretability of machine learning models, which are essential for the accuracy of clinical decision-making.

We first extracted and preprocessed the patient data to address the above problems of data imbalance and feature selection. The preprocessing includes feature vectorization of medical data, missing data processing, normalization, and treatment of class imbalance problems. SMOTE addresses the class imbalance in the target variables, improving our model's accuracy (e.g., Fig. 6) and confirming the importance of data preprocessing techniques for enhancing predictive results in medical research. The feature selection process is performed by arranging the features in order of importance and iteratively eliminating the features that cause a decrease in the AUC value. This leads to an improved combination of predictors, highlighting the importance of feature selection in augmenting model predictions (e.g., Fig. 5).

We chose LNN as the main model for this study because of its ability to process data sequentially and adapt to data changes in real-time<sup>32,33</sup>. This mechanism is similar to how the human brain works, giving LNNs a significant advantage when dealing with medical prediction problems. First, the dynamic architecture of LNN improves the interpretability of the model, and clinical decisions need to be based precisely on understandable and interpretable model results. Second, LNN's continuous learning and adaptation enables it to adapt to changing data after training, which is particularly important for the ability to generalize to small sample data. In addition, LNN has a lower structural and parameter complexity than traditional deep learning models. This makes it less prone to overfitting when trained on small sample datasets while reducing the difficulty of model training and implementation. The LNN model presented in this paper uses only 27 features, and its performance is better than the other four models. In particular, its number of features is also less than that of the LR, RF, and XGBoost models.

SHAP analysis provided insight into the LNN model in this study, revealing the features of greatest importance for predicting AP severity. Among the key factors, Ca level, AMY activity, BAS%,  $CO_2CP$  and EOS% were identified as the main predictors of AP severity. Abnormal Ca levels were strongly associated with the severity of AP, and hypocalcemia was a predictor of intensive care admission or death<sup>34</sup>. Elevated activity of AMY, a biomarker of pancreatitis, is associated with an increased risk of developing AP<sup>35</sup>. Basophils play an essential role in the immune response by releasing inflammatory mediators such as histamine and leukotrienes. The activity and number of basophils increase significantly in pathological states such as allergic reactions and parasitic infections. In addition, basophils are involved in inflammatory processes and influence the function of other immune cells by secreting cytokines and chemokines<sup>36</sup>. In autoimmune pancreatitis (AIP) studies, infiltration of basophils has been found to be more common in pancreatic tissue samples from patients with AIP. Moreover, basophils are activated through the Toll-like receptor (TLR) signaling pathway and play a role in the pathophysiological process of AIP<sup>37</sup>. In some inflammatory diseases, basophil levels are strongly correlated with disease severity. For example, peripheral blood basophil levels are positively correlated with disease severity scores in chronic sinusitis. This suggests that basophils contribute to the progression of the inflammatory response<sup>38</sup>. Thus, it is considered that basophils may, on the one hand, exacerbate tissue damage in the inflammatory response of AP by releasing various inflammatory mediators; on the other hand, they may regulate the activity of other immune cells (e.g., neutrophils, monocytes) through the secretion of cytokines and chemokines, which may affect the inflammatory



response of AP. Eosinophils exert a major impact on the inflammatory response, participating in tissue injury and inflammation by releasing a variety of cytokines, chemokines, and toxic proteins (e.g., eosinophil cationic proteins, major basic proteins, etc.)<sup>39,40</sup>. The inflammatory response is one of the central mechanisms underlying the AP pathophysiologic process, and eosinophil activation and inflammatory mediator release may exacerbate pancreatic tissue damage. Eosinophils also perform an essential function in immune regulation, as they can affect the immune response through interactions with T cells, B cells, and other immune cells<sup>41</sup>. In the pathology of AP, hyperactivation of the immune system may lead to SIRS, and eosinophils may influence the severity and regression of AP by modulating the immune response. AP patients often suffer from metabolic disorders due to inflammatory reactions, tissue necrosis, infections, etc., resulting in an imbalance of acid-base balance. Changes in  $CO_2CP$  can reflect the state of acid-base balance in the patient's body, helping to determine the severity of the condition and guide treatment<sup>42,43</sup>. Early serum  $\alpha$ -HBDH in AP is closely associated with poor prognosis in AP patients. It can be used as a potential early biomarker to predict the severity of AP<sup>44</sup>. In addition, A/G, HDL-C, TG, and CRP were identified as significant predictors of inflammation, lipid metabolism, and liver function, which exert vital roles in the pathologic process of AP<sup>45–48</sup>. These findings validate known AP-related factors and reveal new potential predictors, which provide a more comprehensive perspective on predicting AP severity and further confirm the possible application of the LNN model in medical prediction.

This study still has some limitations as well. First, our study was retrospective, which restricts the ability to assess the model's capacity to predict SAP performance in prospective clinical studies. In addition, this study failed to fully utilize biomarker data such as trypsinogen activating peptide (TAP)<sup>49</sup> and serum macrophage migration inhibitory factor (MIF)<sup>50</sup>. These data are not collected nationally as a standard of care. They are not readily available from electronic medical records, making it impossible to assess the model's performance in identifying SAP cases using this alternative gold standard. Further, hyperparameter tuning is an essential step in developing a machine-learning model and may significantly affect the model's performance. Due to time and resource constraints, this study could not conduct an exhaustive search and test all possible hyperparameter combinations, which may have limited the maximization of model performance. Finally, our data are single-center data, which may lead to bias and limit the generalizability of the model. We will follow up with multiple hospitals by incorporating data from more centers to further validate and optimize our model.

## Conclusions

The LNN-based AP severity prediction model developed in this study is superior to the traditional machine learning model in terms of performance, with the best AUC value and other key statistical indicators. The validity of the LNN model is validated, showing its potential in the clinical prediction of AP severity. This study provides strong scientific support for clinical decision-making through the SMOTE technique and feature selection optimization, combined with SHAP analysis to identify key predictors.

## Data availability

The datasets and codes for this study can be found in the <https://github.com/longshike/LNN-for-SAP-Prediction>.

Received: 4 February 2025; Accepted: 5 May 2025

Published online: 13 May 2025

## References

1. Tenner, S., Baillie, J., DeWitt, J. & Vege, S. S. American college of gastroenterology guideline: Management of acute pancreatitis. *Off. J. Am. Coll. Gastroenterol. ACG* **108**, 1400–1415 (2013).
2. Sinonquel, P., Laleman, W. & Wilmer, A. Advances in acute pancreatitis. *Curr. Opin. Crit. Care* **27**, 193–200 (2021).
3. Na Zhang, X. G., Haiyan, Zhang & Liu, L. Meta-analysis of the etiologic changes of acute pancreatitis in china in the last decade. *Chin. J. Digest. Dis. Imaging: Electron. Ed.* **6**, 71–75 (2016).
4. Khan, J. A. Role of crp in monitoring of acute pancreatitis. *Clinical Significance of C-reactive Protein*, 117–173 (2020).
5. JH, R. Prognostic signs and the role of operative management in acute pancreatitis. *Surg. Gynecol. Obstet.* **139**, 69–81 (1974).
6. Valverde-López, F. et al. Bisap, ranson, lactate and others biomarkers in prediction of severe acute pancreatitis in a European cohort. *J. Gastroenterol. Hepatol.* **32**, 1649–1656 (2017).
7. Wu, B. U. et al. The early prediction of mortality in acute pancreatitis: A large population-based study. *Gut* **57**, 1698–1703 (2008).
8. Hagjer, S. & Kumar, N. Evaluation of the bisap scoring system in prognostication of acute pancreatitis—A prospective observational study. *Int. J. Surg.* **54**, 76–81 (2018).
9. Fan, D. *Research on the role of acute gastrointestinal injury score in assessing the severity and prognosis of acute pancreatitis*. Master's thesis, People's Liberation Army Naval Medical University of China (2021).
10. Wan, J. et al. Serum creatinine level and apache-ii score within 24 h of admission are effective for predicting persistent organ failure in acute pancreatitis. *Gastroenterol. Res. Pract.* **2019**, 8201096 (2019).
11. Thandassery, R. B. et al. Hypotension in the first week of acute pancreatitis and apache ii score predict development of infected pancreatic necrosis. *Dig. Dis. Sci.* **60**, 537–542 (2015).
12. Minyue, Y. et al. Research progress in machine learning modeling of acute pancreatitis. *J. Clin. Hepatol.* **39** (2023).
13. Wu, B. U. et al. Dynamic measurement of disease activity in acute pancreatitis: The pancreatitis activity scoring system. *Off. J. Am. Coll. Gastroenterol. ACG* **112**, 1144–1152 (2017).
14. Han, T. et al. Development and validation of a novel prognostic score based on thrombotic and inflammatory biomarkers for predicting 28-day adverse outcomes in patients with acute pancreatitis. *J. Inflamm. Res.* **15**, 395–408 (2022).
15. Xu, F. et al. Prediction of multiple organ failure complicated by moderately severe or severe acute pancreatitis based on machine learning: A multicenter cohort study. *Mediat. Inflamm.* **2021**, 5525118 (2021).
16. Ye, J.-F., Zhao, Y.-X., Ju, J. & Wang, W. Building and verifying a severity prediction model of acute pancreatitis (ap) based on bisap, mews and routine test indexes. *Clin. Res. Hepatol. Gastroenterol.* **41**, 585–591 (2017).
17. Zhao, B. et al. Cardiac indicator ck-mb might be a predictive marker for severity and organ failure development of acute pancreatitis. *Ann. Transl. Med.* **9**, 368 (2021).

18. Pearce, C. B., Gunn, S. R., Ahmed, A. & Johnson, C. D. Machine learning can improve prediction of severity in acute pancreatitis using admission values of apache ii score and c-reactive protein. *Pancreatology* **6**, 123–131 (2006).
19. Guo, X. *Artificial Intelligence in Predicting the Severity of Acute Pancreatitis*. Master's thesis, Zhejiang University (2020).
20. Sun, H.-W. et al. Accurate prediction of acute pancreatitis severity with integrative blood molecular measurements. *Aging (Albany NY)* **13**, 8817 (2021).
21. Jin, X. et al. Comparison of mpl-ann and pls-da models for predicting the severity of patients with acute pancreatitis: An exploratory study. *Am. J. Emerg. Med.* **44**, 85–91 (2021).
22. Žorňak, M., Beyer, G. & Mayerle, J. Risk stratification and early conservative treatment of acute pancreatitis. *Visceral Med.* **35**, 82–89 (2019).
23. Siregar, G. A. & Siregar, G. P. Management of severe acute pancreatitis. *Open Access Maced. J. Med. Sci.* **7**, 3319 (2019).
24. Thapa, R. et al. Early prediction of severe acute pancreatitis using machine learning. *Pancreatology* **22**, 43–50 (2022).
25. Kui, B. et al. Easy-app: An artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis. *Clin. Transl. Med.* **12**, e842 (2022).
26. Hong, W. et al. Usefulness of random forest algorithm in predicting severe acute pancreatitis. *Front. Cell. Infect. Microbiol.* **12**, 893294 (2022).
27. Yin, M. et al. Automated machine learning for the early prediction of the severity of acute pancreatitis in hospitals. *Front. Cell. Infect. Microbiol.* **12**, 886935 (2022).
28. Yuan, L. et al. Machine learning model identifies aggressive acute pancreatitis within 48 h of admission: A large retrospective study. *BMC Med. Inform. Decis. Mak.* **22**, 312 (2022).
29. Zhou, Y. et al. Prediction of the severity of acute pancreatitis using machine learning models. *Postgrad. Med.* **134**, 703–710 (2022).
30. Luo, Z. et al. Development and evaluation of machine learning models and nomogram for the prediction of severe acute pancreatitis. *J. Gastroenterol. Hepatol.* **38**, 468–475 (2023).
31. Ma, T. *Using Machine Learning to Predict the Severity of Acute Pancreatitis*. Ph.D. thesis, ResearchSpace@ Auckland (2023).
32. Hasani, R., Lechner, M., Amini, A., Rus, D. & Grosu, R. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 7657–7666 (2021).
33. Hasani, R. et al. Closed-form continuous-time neural networks. *Nat. Mach. Intell.* **4**, 992–1003 (2022).
34. Yan, T. et al. Calcium administration appears not to benefit acute pancreatitis patients with hypocalcemia. *J. Hepatobiliary Pancreat. Sci.* **31**, 273–283 (2024).
35. Kim, K. et al. Evaluation of the clinical usefulness of pancreatic alpha amylase as a novel biomarker in dogs with acute pancreatitis: A pilot study. *Vet. Q.* **44**, 1–7 (2024).
36. Miyake, K., Ito, J. & Karasuyama, H. Role of basophils in a broad spectrum of disorders. *Front. Immunol.* **13**, 902494 (2022).
37. Yanagawa, M. et al. Correction to: Basophils activated via tlr signaling may contribute to pathophysiology of type 1 autoimmune pancreatitis. *J. Gastroenterol.* **53**, 582 (2018).
38. Fan, Y., Jiao, Q., Zhou, A. & Liu, J. Correlation between chronic sinusitis subtypes and basophil levels in peripheral blood. *J. Clin. Otorhinolaryngol. Head Neck Surg.* **37**, 293–301 (2023).
39. Matucci, A. et al. Baseline eosinophil count as a potential clinical biomarker for clinical complexity in egpa: A real-life experience. *Biomedicine* **10**, 2688 (2022).
40. Constantine, G. M. & Klion, A. D. Recent advances in understanding the role of eosinophils. *Fac. Rev.* **11**, 26 (2022).
41. Valent, P. et al. Eosinophils and eosinophil-associated disorders: immunological, clinical, and molecular complexity. In *Seminars in immunopathology*, vol. 43, 423–438 (Springer, 2021).
42. Achanti, A. & Szerlip, H. M. Acid-base disorders in the critically ill patient. *Clin. J. Am. Soc. Nephrol.* **18**, 102–112 (2023).
43. Seifter, J. L. & Chang, H.-Y. Disorders of acid-base balance: New perspectives. *Kidney Dis.* **2**, 170–186 (2017).
44. Xiao, W. et al. Serum hydroxybutyrate dehydrogenase as an early predictive marker of the severity of acute pancreatitis: A retrospective study. *BMC Gastroenterol.* **20**, 1–7 (2020).
45. Yuan, L. et al. Machine learning model identifies aggressive acute pancreatitis within 48 h of admission: A large retrospective study. *BMC Med. Inform. Decis. Mak.* **22**, 312 (2022).
46. Krasztel, M. M. et al. Accuracy of acute-phase proteins in identifying lethargic and anorectic cats with increased serum feline pancreatic lipase immunoreactivity. *Vet. Clin. Pathol.* **51**, 93–100 (2022).
47. Wang, B. et al. Lipid levels and risk of acute pancreatitis using bidirectional mendelian randomization. *Sci. Rep.* **14**, 6267 (2024).
48. Mofidi, R. et al. Association between early systemic inflammatory response, severity of multiorgan dysfunction and death in acute pancreatitis. *J. Br. Surg.* **93**, 738–744 (2006).
49. Shaheen, Y. A. R., Aglan, B. M., Okab, E. A. & El Gouhary, A. S. Comparison of clinical outcomes between aggressive and non-aggressive intravenous hydration for pancreatitis. *Benha J. Appl. Sci.* **9**, 121–137 (2024).
50. Shen, D., Tang, C., Zhu, S. & Huang, G. Macrophage migration inhibitory factor is an early marker of severe acute pancreatitis based on the revised Atlanta classification. *BMC Gastroenterol.* **21**, 1–8 (2021).

## Acknowledgements

We would like to thank all the participants in this study for their contributions and the Gastroenterological Diseases Branch of Guangxi Medical Association for their guidance on our manuscript.

## Author contributions

Jie Cao, Shike Long and Liu Ying contributed in the conception and design. Jie Cao, Huan Liu, Fu'an Chen, Shiwei Liang and Haicheng Fang contributed in the collection and assembly of data. Jie Cao, Shike Long and Ying Liu contributed in the development of methodology. Jie Cao, Shike Long, Huan Liu, Liu Ying and Fu'an Chen contributed in the data analysis and interpretation. Jie Cao and Shike Long contributed in the manuscript drafting. Cao Jie, Shike Long and Ying Liu contributed in the manuscript revision. All authors read and approved the final manuscript.

## Funding

This work was supported by the Central Guided Local Science and Technology Development Fund Project (No. Guike ZY23055033) the Guangxi Natural Science Foundation (No. 2024GXNSFAA010088) the Guilin Scientific Research and Technology Development Program Project (No. 20230116-2) the Guangxi medical and health care appropriate technology development and popularization and application project (No. S2023126) and the Guangxi Medical and Health Key Cultivation Discipline Construction Project.

## Declarations

### Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

### Ethics approval and consent to participate

Studies involving human participants underwent a review and approval process by the Ethics Committee of the Second Affiliated Hospital of Guilin Medical University. Patients participating in the study signed an informed consent form through themselves and/or their legal guardians. The study was approved by the Ethics Committee of the Second Affiliated Hospital of Guilin Medical University (NO.ZLXM-2024013).

### Additional information

**Correspondence** and requests for materials should be addressed to Y.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025