Scientific Reports

https://doi.org/10.1038/s41598-026-38410-0

Article in Press

Predicting infected pancreatic necrosis in acute pancreatitis using machine learning models and feature selection

Received: 23 May 2025

Accepted: 29 January 2026

Published online: 28 February 2026

Cite this article as: Xin L., Yixuan D., Bohan H. *et al.* Predicting infected pancreatic necrosis in acute pancreatitis using machine learning models and feature selection. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38410-0

Li Xin, Ding Yixuan, Huang Bohan, Shen Yunheng, Lv Hairong, Cao Feng, Yu Tong, Li Fei, Fei Xiaolu & Li Jia

Predicting Infected Pancreatic Necrosis in Acute Pancreatitis Using Machine Learning Models and Feature Selection

Li Xin[1,†], Ding Yixuan[3,†], Huang Bohan[1], Shen Yunheng[2], Lv Hairong[2], Cao Feng[3], Yu Tong[4], Li Fei[3,*], Fei Xiaolu[3,*], Li Jia[3,4,*]

[†]Li Xin and Ding Yixuan contributed equally to this work.

[*]Correspondence:lifei@xwhospital.com; feixiaolu@xwhospital.com; lijia@xwh.ccmu.edu.cn

Full list of author information is available at the end of the article.

Abstract

Background: Infected pancreatic necrosis (IPN) is a life-threatening complication of acute pancreatitis (AP), and its early prediction remains challenging. This study aimed to develop and externally validate interpretable machine learning models for individualized IPN risk prediction.

Methods: A total of 728 patients with AP admitted to Xuanwu Hospital, Capital Medical University, between 2017 and 2023 were retrospectively analyzed. Embedded feature selection was incorporated within model training using regularized linear and tree-based algorithms to enhance interpretability and prevent overfitting. Five machine learning algorithms and one neural network model were evaluated through nested cross-validation and an independent temporal external cohort consisting of 166 AP patients admitted to Xuanwu Hospital, Capital Medical University, between 2022 and 2023. Model discrimination, precision–recall, and probability calibration were assessed, and model explainability was analyzed using Shapley Additive Explanations (SHAP).

Results: The Random Forest model achieved the best overall performance, achieving an external AUC of 0.764 (95% CI: 0.696–0.830, $P < 0.001$), precision of 0.893, recall of 0.604, and the lowest Brier score, indicating reliable probability calibration. SHAP analysis identified Fibrinogen, APACHE II score, D-dimer, IL-6, and C-reactive protein (CRP) as key predictors associated with increased IPN risk, while higher Lymphocyte count, and Hematocrit were protective. These findings are consistent clinical pathophysiology.

Conclusion: The interpretable Random Forest model demonstrated robust discrimination and calibration for IPN prediction, providing a transparent and data-driven framework for early risk stratification in acute pancreatitis. Prospective multicenter validation is warranted before clinical implementation.

Keywords

Acute pancreatitis ; Deep learning ; Infected pancreatic necrosis ; Machine learning ; Prognosis prediction

Introduction

Acute pancreatitis (AP) is a prevalent acute digestive system disorder characterized by its high incidence and potential for severe outcomes. Globally, the incidence of AP ranges from 13 to 45 cases per 100,000 individuals annually, with a mortality rate of approximately 1%-3% in mild cases and up to 20%-30% in severe cases [1, 2].According to the modified Atlanta classification, AP is categorized into mild, moderate, and severe based on the presence of organ dysfunction and local complications, such as acute pancreatic pseudocysts, fluid collections, and infected pancreatic necrosis (IPN) [3]. Among these complications, IPN is particularly critical, occurring in 30%-40% of patients with severe AP and associated with a mortality rate of 10%-20% [4, 5]. Therefore, early identification of high-risk IPN patients is crucial for timely intervention and improved prognosis.

In the clinical management of AP, various scoring systems and biomarkers have been developed to predict disease severity. Traditional scoring systems, such as the Ranson score [6],

Bedside Index for Severity in Acute Pancreatitis (BISAP) [7], Modified CT Severity Index (MCTSI) [8], and Acute Physiology and Chronic Health Evaluation II (APACHE II) [9], integrate multiple clinical indicators to provide valuable assessment tools for clinicians. Additionally, simplified scoring systems like the Pancreatitis Activity Scoring System (PASS) [10] and the Chinese Severe Acute Pancreatitis Scoring System (CSSS) [11] have been introduced to enhance the convenience and timeliness of early prediction. However, these systems exhibit significant limitations in accurately predicting IPN, a severe complication of AP. They often lack precision in identifying high-risk patients and fail to account for the unique pathophysiology of IPN, resulting in suboptimal sensitivity and specificity. To address these limitations, machine learning (ML) and deep learning (DL), as branches of artificial intelligence, have emerged as powerful tools for disease prediction.

ML and DL models can integrate diverse clinical variables and improve predictive accuracy compared to traditional methods [12]. In one study, logistic regression (LR), Fully connected neural network (FCNN), and Extreme Gradient Boosting (XGBoost,XGB) were used to predict the risk of severe acute pancreatitis (SAP), demonstrating that ML models outperformed traditional scoring systems in risk stratification[13]. Recent advances show that ML with feature optimization and explainable AI (XAI) can achieve both high performance and transparency across clinical domains. In dengue diagnosis, Recursive Feature Elimination with Shapley Additive Explanations (SHAP) improved Random Forest (RF) performance while identifying key predictors [21]. Similarly, in maternal health, Linear Discriminant Analysis (LDA)-optimized features with ensemble learning achieved high accuracy, with SHAP clarifying decision logic [22].

Evidence from hepatology further supports this approach. Studies using Deep Feature Synthesis with tree ensembles achieved strong performance and provided interpretability through SHAP analysis [23]. Other work combined tree-based selection with ensemble models and XAI, though noting limitations like dataset bias and lack of external validation [25]. An Artificial Neural Network (ANN)-based framework optimized by LDA significantly improved accuracy while systematically integrating SHAP, Local Interpretable Model-agnostic Explanations (LIME), and Individual Conditional Expectation (ICE), emphasizing the need for rigorous val-

idation [24].

Although ML has been increasingly applied in predicting disease severity, previous studies on AP mainly focused on overall severity rather than IPN, and often lacked external validation or model interpretability. To address these gaps, this study developed and externally validated an interpretable ML framework for individualized IPN prediction. Multiple algorithms, including linear, tree-based, and neural network models, were systematically compared under nested cross-validation and temporal validation. Embedded feature selection methods (L1, Elastic Net, and tree-based importance) and SHAP interpretation were integrated to enhance model robustness and clinical explainability.

This approach combines methodological rigor and interpretability, distinguishing our work from previous studies and providing a reliable tool for early IPN risk assessment.

Method

Data Collection and Processing

This was a single-center retrospective study conducted at Xuanwu Hospital, Capital Medical University, including 728 patients with AP admitted between December 2017 and December 2023. Patients included in this study were 18 years of age or older and were diagnosed with a first episode of AP according to the 2012 Atlanta Classification criteria. The classification criteria for AP refer to the *Revised Atlanta Classification and Definitions by International Consensus* [3]. Exclusion criteria included: (1) lack of early disease information, no abdominal computed tomography (CT) scan within one week of onset, or no serum marker tests within 24 hours of onset; (2) patients with recurrent AP; (3) patients with chronic pancreatitis, pancreatitis related to trauma or pregnancy; (4) patients with cancer; (5) patients under 18 years of age. After applying these criteria, 432 patients were classified into the IPN group, and 296 into the non-IPN group. A flow diagram of the selection process is presented in Supplementary Figure S3.

Most predictor variables were collected within 24 hours of admission. Additional variables—such as comorbidities, MCTSI, and organ failure—were obtained at predefined time

points but always before IPN diagnosis. Since IPN typically developed days after admission, all variables were available prior to outcome determination. Details are provided in Supplementary Table S3.

The primary outcome was IPN, defined according to the 2012 Revised Atlanta Classification and supported by the latest international consensus guidelines [3, 26]. Diagnosis required evidence of pancreatic or peripancreatic necrosis with either (1) microbiological confirmation or (2) gas within the necrotic area on contrast-enhanced computed tomography. The diagnosis date was defined as the earliest time either criterion was met.

This study was approved by the Ethics Committee of Xuanwu Hospital, Capital Medical University (Approval No.: XA Lin Yan Shen [KS2025] 002-001). All analyses were performed using anonymized data. Due to the retrospective nature of the study and the use of de-identified records, the requirement for individual informed consent was waived. All procedures were conducted in accordance with the principles of the Declaration of Helsinki.

In this study, we preprocessed the raw data to ensure its suitability for training and prediction with ML and FCNN. The main steps of preprocessing include binarization, one-hot encoding, normalization, and data distribution analysis.

Missing values were imputed using the median (for continuous variables) or mode (for categorical variables). The proportion of missing data for each variable is provided in Supplementary Table S4. To prevent data leakage, all preprocessing steps, including imputation and normalization, were fitted exclusively on the training folds during cross-validation and subsequently applied to the corresponding validation folds and the external test set. To address class imbalance, inverse class weighting was incorporated into the loss function, and stratified sampling was employed during cross-validation to ensure representative data splits. Categorical variables were one-hot encoded prior to model training, expanding several variables into multiple dummy features and resulting in a maximum of 33 input dimensions. Feature normalization was performed according to the following formula:

$$data[i]_{\text{norm}} = \frac{data[i] - data_{\min}}{data_{\max} - data_{\min}} \tag{1}$$

where $data[i]$ is the i-th data point, and $data_{\min}$ and $data_{\max}$ are the minimum and maximum

values, respectively.

All data preprocessing and model development were conducted in Python (v3.10) on a Windows-based environment using commonly adopted scientific computing and machine-learning libraries, including NumPy, pandas, scikit-learn, XGBoost, LightGBM, and Matplotlib/Seaborn for visualization; SHAP analyses were performed using the SHAP Python package. All code was executed in isolated environments to ensure reproducibility, and the exact package versions were recorded programmatically and are provided as supplementary material or are available upon reasonable request. All stochastic procedures used a fixed random seed, which is documented in the analysis scripts and available upon reasonable request.

Model Development and Validation

Six predictive models were developed: gradient boosting machine (GBM), XGB, RF, support vector machine (SVM), LR, and a FCNN. The data were temporally split, with samples from 2017 to 2021 used for model development, and 2022 to 2023 held out as an independent external test set for temporal validation. Nested cross-validation was applied within the training period: the outer 5-fold CV was used to estimate model performance, while the inner 3-fold CV was used for hyperparameter tuning and feature selection.

To mitigate data leakage, embedded feature selection was integrated directly into the model training process within a nested cross-validation framework applied exclusively to the training set (2017–2021). Within this framework, hyperparameters were optimized and feature subsets were determined independently inside each training fold.

The final model was subsequently constructed by applying the optimized training procedure to the entire training set, and its generalizability was evaluated through a single assessment on the held-out temporal test set (2022–2023).

All models were implemented using scikit-learn and XGBoost, with class weighting applied to address class imbalance. Hyperparameters were tuned using 5-fold inner cross-validation, optimizing the area under the receiver operating characteristic curve (ROC-AUC).

For linear models, embedded feature selection was performed via L1 or Elastic Net regularization in combination with SelectFromModel, retaining features with non-zero coefficients

Table 1: Clinical characteristics of patients (n = 728)

	IPN (n=432)	Non-IPN (n=296)	P value
Epidemiology			
Age, median (min, max), y	51.03 (11-90)	52.14 (14-99)	0.374
BMI, median (min, max),kg/m^2	23.89 (15.63-41.52)	24.62 (15.05-39.33)	0.024
Gender, n (%)			0.62
Male	282 (65.3%)	199 (67.2%)	
Female	150 (34.7%)	97 (32.8%)	
Medical History, n (%)			
Cardiovascular	178 (41.2%)	120 (40.5%)	
Cerebrovascular	25 (5.8%)	16 (5.4%)	
Respiratory	12 (2.8%)	11 (3.7%)	
Nephropathy	16 (3.7%)	9 (3.0%)	
Diabetes	78 (18.1%)	54 (18.2%)	
Substance Abuse, n (%)			
Smoking History	119 (27.5%)	99 (33.4%)	
Alcohol History	130 (30.1%)	101 (34.1%)	
Laboratory, Median (min, max)			
APACHE II	12.98 (1-36)	11.84 (1-34)	0.012
MCTSI	8.49 (2-10)	6.49 (2-10)	0.805
WBC (10^9/L)	11.13 (2.07-33.15)	11.70 (2.8-54.01)	0.230
Hematocrit	31.07 (12.2-55.9)	37.49 (16.4-57.3)	0.436
Lymph (10^9/L)	1.19 (0.07-15.94)	1.10 (0.14-3.29)	0.136
PLT (10^9/L)	282.13 (20-800)	247.67 (33-1084)	$<$0.001
Neut (10^9/L)	9.27 (1.21-30.23)	9.95 (1.17-51.33)	0.132
International Normalized Ratio	1.27 (0.9-11.93)	1.16 (0.87-2.52)	0.002
Activated Partial Thromboplastin Time (s)	42.93 (0-98.6)	40.43 (22.7-101.1)	$<$0.001
Thrombin Time (s)	16.17 (0-124.7)	16.12 (0-44.5)	0.897
Prothrombin Time (s)	15.79 (12.2-94.2)	14.81 (12-27.5)	$<$0.001
Fibrinogen (g/L)	4.92 (0.6-13.71)	5.28 (1.12-15)	0.011
D-Dimer (μg/mL)	4.55 (0-19.53)	4.51 (0-19.11)	0.909
Creatinine (μmol/L)	89.63 (10-953)	89.23 (15-987)	0.961
C-Reactive Protein (mg/L)	136.59 (2-571)	157.76 (3-471)	0.004
White Blood Interleukin-6 (pg/ml)	194.01 (0-5000)	242.15 (0-3666)	0.129
Procalcitonin (ng/ml)	4.03 (0-100)	4.17 (0-83.9)	0.871
Etiology			
Biliary	240 (55.6%)	156 (52.7%)	
Hypertriglyceridemia	127 (29.4%)	91 (30.7%)	
Alcoholic	18 (4.2%)	17 (5.7%)	
Other	47 (10.9%)	32 (10.8%)	
Complications, n (%)			
Respiratory Failure	122 (28.2%)	54 (18.2%)	
Renal Failure	81 (18.8%)	24 (8.1%)	
Heart Failure	91 (21.1%)	24 (8.1%)	

Data are presented as median (minimum–maximum) or count (percentage). Continuous variables were compared using the Mann–Whitney U test, and categorical variables were compared using the chi-square or Fisher's exact test as appropriate. P values < 0.05 were considered statistically significant. APACHE II, Acute Physiology and Chronic Health Evaluation II; MCTSI, Modified Computed Tomography Severity Index; WBC, White Blood Cell Count; Lymph, Lymphocyte Count; PLT, Platelet Count; Neut, Neutrophil Count.

within each training fold to prevent information leakage. The FCNN employed L1 or Elastic Net regularization at the input layer to encourage sparsity, depending on the configuration.

For tree-based models (RF, GBM, and XGBoost), no explicit feature elimination was conducted; instead, features were implicitly weighted according to split gain or impurity reduction during training. Importantly, all feature selection or weighting processes were confined strictly to the training folds.

The FCNN architecture consisted of one or two hidden layers containing 64 to 128 neurons per layer, with activation functions chosen from ReLU and Tanh. We performed hyperparameter optimization through a grid search coupled with five-fold cross-validation. The search employed the Adam optimizer with learning rates ranging from 0.0005 to 0.001 and incorporated early stopping (patience=10). A summary of the hyperparameter search space and the resulting optimal configurations are provided in Table 2 and Supplementary Table S1, respectively.

Table 2: Architecture and hyperparameter search space of the FCNN.

Parameter	**Range / Setting**	**Description**
Input dimension	12, 15, 20, 31	Based on feature configuration
Hidden layers	(64,), (64, 32)	One or two layers
Activation function	ReLU, Tanh	Nonlinear transformation
Optimizer	Adam	Adaptive gradient optimization
Learning rate	0.0005, 0.001	Grid search range
Regularization (α)	0.001, 0.01, 0.1	L2 penalty
Max iterations	3000	Early stopping enabled
Patience	10 epochs	Validation monitored

The FCNN was tuned using a grid search within nested cross-validation, where hyperparameters were optimized based on ROC-AUC performance in the inner folds. Early stopping with a patience of 10 epochs was applied to prevent overfitting. The learning rate and regularization parameter (α) were selected according to validation performance. ReLU and Tanh activations were tested to balance convergence speed and nonlinearity.

Model performance was primarily assessed using ROC-AUC. Secondary metrics included accuracy, precision, recall, F1-score, and Brier score. Ninety-five percent confidence intervals (95% CIs) for ROC-AUC, accuracy, and F1-score were estimated via 1000 bootstrap iterations. Calibration and precision–recall (PR) curves were plotted for visual assessment. The decision threshold for binary classification was determined by Youden's Index, which maximizes the sum of sensitivity and specificity on the ROC curve. Specifically, the threshold was selected at the point where the true positive rate and the true negative rate were maximized. Pairwise ROC-

AUC differences between models were evaluated using a bootstrap-based DeLong-like test and visualized as heatmaps. The best-performing model was further interpreted using SHAP values to quantify feature contributions.

Result

Model Performance and Feature Selection

All models were trained using embedded feature selection within a nested cross-validation framework, resulting in a spectrum of model complexities defined by the retention of 5 to 33 features. Tree-based ensembles—RF, GBM, and XGB—achieved the strongest external performance, with ROC-AUC values approaching 0.76. The FCNN with elastic-net regularization followed closely, achieving an ROC-AUC of 0.74, while linear models and the SVM delivered slightly lower results, with ROC-AUCs in the range of 0.68 to 0.73. The robust generalizability of all models was further confirmed by internal cross-validation, which yielded consistently high mean ROC-AUCs between 0.81 and 0.82.

As a conventional clinical benchmark, the APACHE II score was also evaluated independently on the external validation set. Its predictive performance was notably lower than that of most ML models, with a ROC-AUC of 0.596 and a PR-AUC of 0.695, highlighting the potential added value of multivariable ML approaches for IPN prediction.

Figure 1 displays the Receiver Operating Characteristic (ROC) and PR curves across all ML models, while the PR curve of the APACHE II score is provided separately in the Supplementary Materials. Comparative ROC-AUC and PR-AUC results are summarized in Figure 2. The primary performance indicators, including ROC-AUC, accuracy, and F1-score, are reported in Table 3, while additional metrics such as precision, recall, Brier score, and classification thresholds are summarized in Supplementary Table S2.

Statistical and Calibration Analyses

Supplementary Figure S1 shows the results of statistical comparison and calibration analysis. Subplots (a)–(c) present pairwise ROC-AUC comparisons using DeLong tests under different

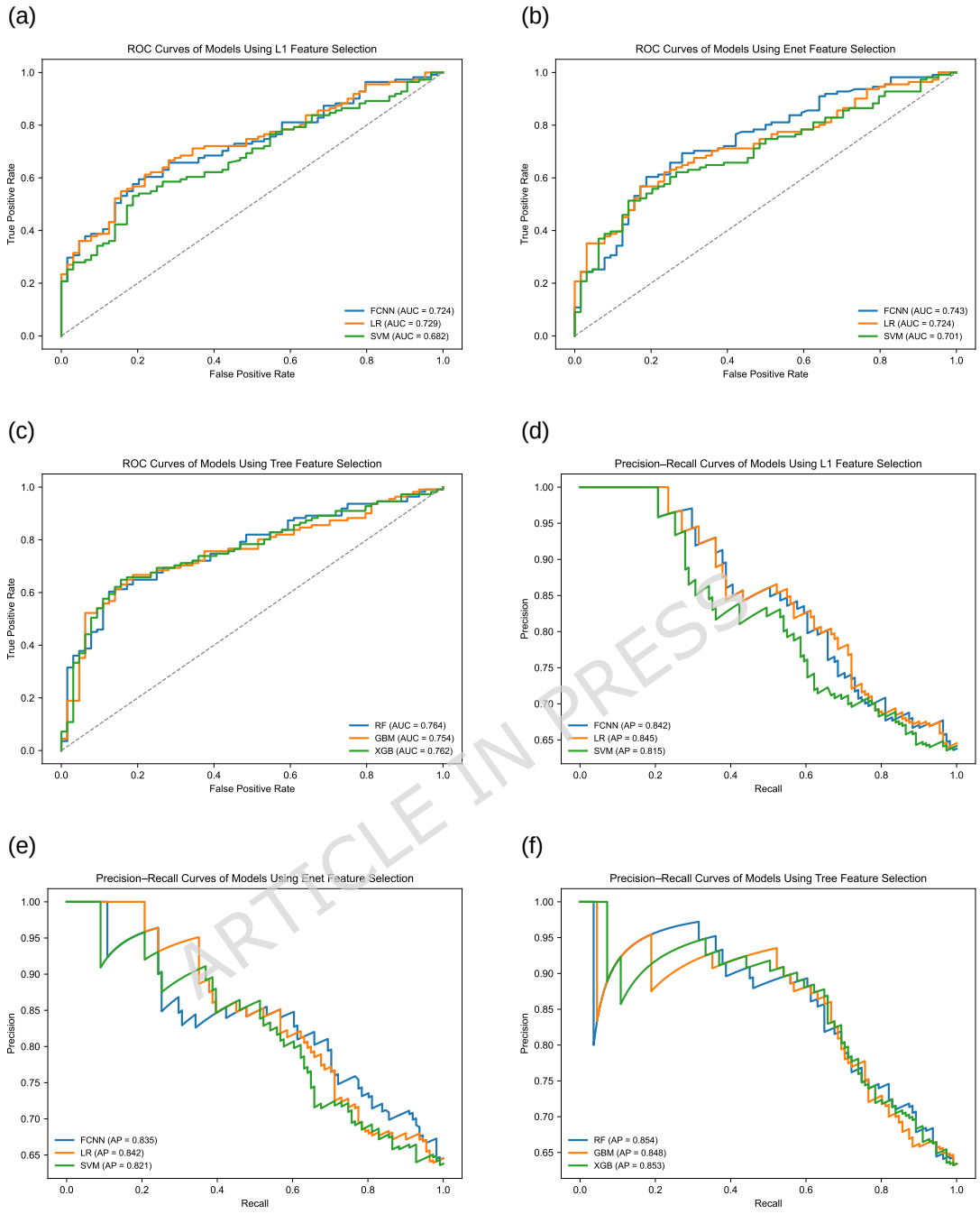

Figure 1: **Model performance comparison by ROC-AUC and PR-AUC in the external validation cohort.** Each panel illustrates the Receiver ROC and PR curves for models developed under three feature selection strategies: L1 regularization, Elastic Net, and tree-based embedded importance. Shaded areas represent 95% confidence intervals derived from 1000 bootstrap resamples. The decision threshold (marked by solid dots) corresponds to the Youden's Index, maximizing the sum of sensitivity and specificity. AUC values were compared across models using the bootstrap-based DeLong test.

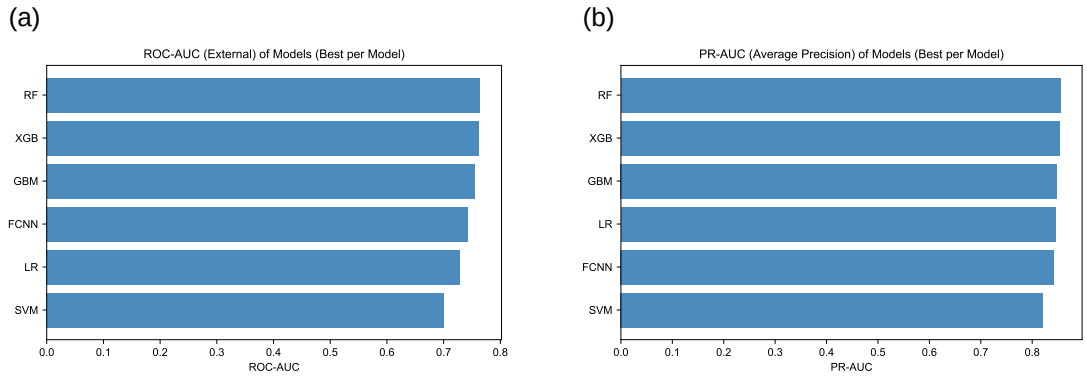

Figure 2: **ROC-AUC and PR-AUC performance across models.** (a) ROC-AUC values of each model on the external validation set. (b) PR-AUC (average precision) comparison across models.

Table 3: Performance metrics of all models on the external validation set.

Model	Features	Method	ROC-AUC	ROC-AUC (95% CI)	Accuracy	F1-score	PR-AUC
FCNN	9	embedded_l1	0.724	(0.653, 0.790)	0.669	0.695	0.842
LR	9	embedded_l1	0.729	(0.657, 0.797)	0.674	0.705	0.845
SVM	33	embedded_l1	0.682	(0.605, 0.755)	0.634	0.648	0.815
FCNN	5	embedded_enet	0.743	(0.664, 0.814)	0.680	0.705	0.835
LR	10	embedded_enet	0.724	(0.655, 0.794)	0.663	0.681	0.842
SVM	17	embedded_enet	0.701	(0.627, 0.775)	0.640	0.644	0.821
RF	33	embedded_tree	0.764	(0.696, 0.830)	0.703	0.720	0.854
GBM	33	embedded_tree	0.754	(0.682, 0.823)	0.720	0.751	0.848
XGB	33	embedded_tree	0.762	(0.687, 0.831)	0.720	0.746	0.853
APACHE II	N/A	scoring system	0.596	(0.507, 0.682)	0.657	0.752	0.695

feature selection methods. Tree-based models (RF, GBM, XGB) showed similar ROC-AUCs. In contrast, differences were observed between tree-based and linear or neural network models. Subplots (d)–(f) show calibration curves for the corresponding models. Tree-based models demonstrated better calibration, with predicted probabilities closer to the ideal line, compared to linear and neural network models.

Feature Interpretation of the Best Model

RF was selected due to its superior external ROC-AUC and well-calibrated performance. SHAP analysis revealed that higher levels of Fibrinogen, Prothrombin Time, APACHE II score, INR, and inflammatory markers such as IL-6 and C-reactive protein (CRP) were associated with an increased risk of IPN, as shown in Figure 3. Conversely, higher lymphocyte counts and lower creatinine or MCTSI values were linked to reduced IPN risk, indicating their protective roles.

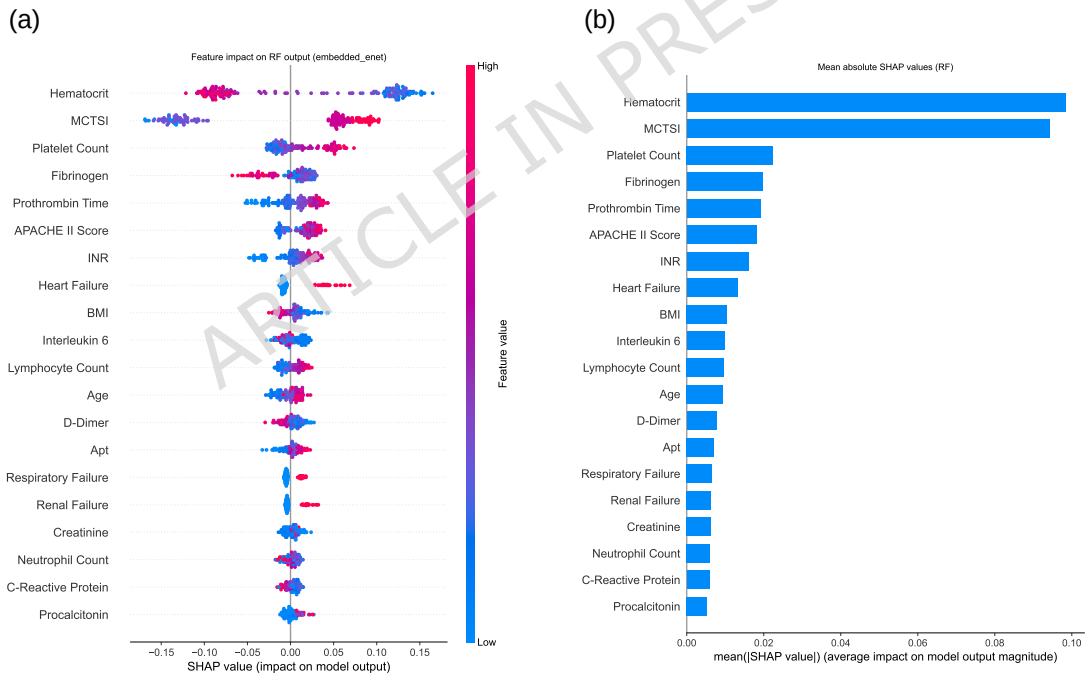

Figure 3: **SHAP-based feature interpretation of the best-performing RF model.** (a) Summary plot showing individual feature contributions to IPN prediction, where red indicates higher and blue lower feature values. (b) Bar plot ranking features by their mean absolute SHAP values, indicating overall feature importance.

Discussion

In this study, we systematically evaluated multiple ML models for predicting IPN using a robust framework that integrated embedded feature selection and temporal external validation. A key finding was the superior and robust performance of tree-based ensemble methods, particularly RF. When trained on features selected by a tree-based selector, RF achieved optimal performance on the independent test set, exhibiting excellent discrimination (AUC = 0.764) and outstanding precision in identifying IPN cases (Average Precision = 0.854). Comparable performance was observed for other tree-based models, including XGBoost and GBM, which were evaluated under the same feature-selection paradigm. In contrast, models relying on L1- or Elastic Net–based feature selection, such as LR, SVM, and FCNN, showed more moderate performance. Collectively, these findings highlight the importance of aligning model architecture with feature selection strategy and suggest that tree-based approaches are particularly effective in capturing the complex, non-linear relationships underlying IPN development.

Earlier prediction models for AP severity and IPN mainly relied on multivariable LR using clinician-selected variables [13, 14]. Although these approaches are transparent and easy to interpret, they often fail to capture nonlinear and high-dimensional interactions in clinical data, which limits their discriminative power and external generalizability.

More recent research has shifted toward machine and deep learning approaches that integrate embedded feature selection, ensemble methods, and XAI to enhance predictive accuracy and interpretability [16, 15]. Clinical studies across various diseases have demonstrated that combining optimized feature selection with XAI techniques—such as SHAP and LIME—can yield models that are both accurate and transparent, supporting more trustworthy decision-making in medical settings [21, 23, 24].

Building on these methodological advances, our study demonstrates the added value of an integrated ML framework for IPN prediction in acute pancreatitis. Among all evaluated models, RF achieved the most favorable external performance, with superior calibration compared to linear and neural network–based approaches. To contextualize the added value of ML, we benchmarked model performance against the conventional APACHE II score. On the temporal test set, APACHE II showed only moderate discrimination (ROC-AUC = 0.596, PR-AUC =

0.695), whereas most ML models achieved ROC-AUC values exceeding 0.70. These results underscore the potential of multivariable learning-based approaches to enhance early IPN risk stratification beyond traditional clinical scoring systems.

Beyond discrimination, calibration is essential for clinical deployment because it determines whether predicted risks correspond to observed outcomes and supports threshold selection tailored to clinical priorities. All tree-based ensembles produced clinically reliable probability estimates, with the RF showing the lowest Brier score of 0.194, as reported in Supplementary Table S2. At the standard threshold, RF also provided high precision (0.893) with moderate recall (0.604), and decision curve analysis further supported its potential clinical utility. In practice, excessive false-positive predictions may lead to unnecessary imaging or antibiotic exposure, whereas false-negative predictions could delay timely intervention. Therefore, well-calibrated risk estimates and flexible threshold adjustment are critical to balancing sensitivity and specificity in real-world use.

Statistical comparisons via pairwise DeLong tests showed no statistically significant differences in discrimination among models within the same feature-selection group, as illustrated in Supplementary Fig. S1(a)–(c), with all $p > 0.05$. Nevertheless, RF maintained a modest numerical advantage over other tree-based ensembles and was further characterized by consistently favorable calibration. These results support selecting RF as the most clinically robust predictor, based on its practical consistency, reliable probability estimates, and interpretability, rather than relying solely on marginal statistical superiority in AUC.

To further enhance model interpretability, SHAP analysis was conducted on the best-performing RF model to quantify each feature's contribution to IPN prediction [17, 18]. The analysis revealed that systemic inflammatory, coagulation, and organ dysfunction markers were the dominant determinants of IPN, as illustrated in Figure 3. Elevated Fibrinogen, Prothrombin Time, INR, CRP, Procalcitonin, Interleukin-6, and D-dimer were associated with increased predicted risk, reflecting the hyperinflammatory and procoagulant state characteristic of severe acute pancreatitis. Conversely, higher Lymphocyte count, and Hematocrit, along with lower Creatinine and MCTSI scores, were linked to reduced risk, suggesting preserved immune competence, adequate perfusion, and less extensive parenchymal necrosis. Established severity indices such

as APACHE II and MCTSI also ranked among the most influential predictors, demonstrating strong concordance between model-derived and clinical understanding.

From a clinical standpoint, the SHAP-derived pattern closely mirrors the pathophysiological cascade leading from sterile necrosis to infection. Elevated coagulation and inflammatory markers indicate endothelial injury and microthrombosis, which compromise pancreatic perfusion and create a favorable environment for bacterial translocation. Meanwhile, lower Hematocrit values reflect hypovolemia and catabolic stress, often observed in patients who progress to IPN. Representative features such as MCTSI, Hematocrit, and D-dimer thus bridge radiologic severity, circulatory disturbance, and coagulation dysfunction—the three interdependent mechanisms underlying infected necrosis. This coherence between model insights and biological mechanisms reinforces both the interpretability and credibility of the predictive framework.

The robustness and generalizability of the models were supported by consistent internal and external performance. During cross-validation, all models achieved mean AUC values between 0.81 and 0.82, with RF and XGB showing particularly stable performance at 0.816 ± 0.020 and 0.817 ± 0.011, respectively. Detailed cross-validation outcomes are provided in Supplementary Table S2. External validation revealed only a minor decline in discrimination, demonstrating good temporal transportability and model stability. These findings indicate that the proposed framework can generalize reliably to new patient populations, despite inherent heterogeneity in disease presentation and clinical practice. However, as a single-center retrospective study, this work remains limited by potential selection bias, residual confounding, and temporal drift. Future multicenter and prospective validation will be essential to confirm its reproducibility and ensure real-world applicability before clinical deployment.

Clinically, the proposed ML framework has the potential to support early risk stratification of IPN, enabling timely identification of high-risk patients and optimization of clinical decision-making. Such predictive tools could guide escalation of care, such as earlier transfer to intensive monitoring, targeted antibiotic therapy, or closer imaging surveillance in patients predicted to develop infection. By integrating routinely available laboratory and clinical parameters, the model is both interpretable and feasible for real-time application in tertiary and emergency care settings.

This study has several limitations that should be acknowledged. First, this study included a moderate sample size (n = 728) with 432 IPN events. As Xuanwu Hospital is a large tertiary referral center, the case mix may be enriched for more severe presentations, which could contribute to a relatively higher IPN proportion compared with primary or community settings. Second, although we implemented nested cross-validation, embedded feature selection, and an independent temporal validation cohort, the possibility of model optimism due to center-specific practice patterns cannot be completely excluded. Importantly, the consistent performance across internal cross-validation and temporal validation suggests that overfitting was mitigated to a substantial extent. Third, while temporal external validation was performed, validation across different institutions is still needed to confirm transportability under varying laboratory assays, clinical workflows, and patient populations. Prospective multicenter studies and, if necessary, model recalibration are warranted before broader clinical implementation.

Future work should prioritize validation in multicenter cohorts spanning different care settings, including both tertiary referral and non-referral hospitals, to further assess transportability and calibration stability across institutions. Model recalibration and, where appropriate, transfer learning approaches may help adapt predictions to evolving clinical standards, laboratory assays, and population characteristics over time. In addition, integrating multimodal data—such as radiological features and longitudinal trajectories of inflammatory markers—may further enhance predictive performance and provide deeper mechanistic insight. Finally, future studies should extend evaluation beyond discrimination and calibration to assess real-world clinical impact, including outcome-oriented and cost-sensitive analyses, to determine whether model-assisted decision-making improves patient outcomes compared with standard practice.

Conclusion

This study developed and evaluated an interpretable ML framework for early prediction of IPN in patients with acute pancreatitis, with IPN defined according to established diagnostic criteria. Predictors were restricted to routinely available variables collected early after admission and prior to IPN diagnosis to minimize information leakage, and embedded feature selection was integrated within a nested cross-validation framework. Using a temporal hold-out cohort for val-

idation, the random forest model achieved the best overall performance, with an external AUC of 0.764, balanced precision and recall, and the lowest Brier score, indicating reliable probability calibration. SHAP analysis highlighted clinically plausible predictors, including fibrinogen, APACHE II score, D-dimer, IL-6, and CRP as risk-enhancing factors, whereas lymphocyte count and hematocrit were associated with reduced risk. These results support the potential of transparent, tree-based models for early, data-driven risk stratification in acute pancreatitis. Future work should prioritize prospective multicenter validation, calibration assessment and recalibration across institutions, and integration of multimodal data to enhance generalizability and real-world applicability.

Abbreviations

AP	Acute pancreatitis
IPN	Infected pancreatic necrosis
BISAP	Bedside Index for Severity in Acute Pancreatitis
MCTSI	Modified Computed Tomography Severity Index
APACHE II	Acute Physiology and Chronic Health Evaluation II
PASS	Pancreatitis Activity Scoring System
CSSS	Chinese Simple Severity Score
ML	Machine learning
DL	Deep learning
LR	Logistic regression
FCNN	Fully connected neural network
XGB	Extreme gradient boosting
SAP	Severe acute pancreatitis
XAI	Explainable artificial intelligence
SHAP	Shapley additive explanations
RF	Random forest
LDA	Linear discriminant analysis
ANN	Artificial neural network
LIME	Local Interpretable Model-agnostic Explanations
ICE	Individual Conditional Expectation
CT	Computed tomography
GBM	Gradient boosting machine
SVM	Support vector machine
ROC-AUC	Area under the receiver operating characteristic curve
PR	Precision–recall
ROC	Receiver operating characteristic
Neut	Neutrophil count
INR	International normalized ratio

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of Xuanwu Hospital, Capital Medical University (Approval No.: XA Lin Yan Shen [KS2025] 002-001). The study used pre-existing, fully anonymized clinical data; therefore, the requirement for individual informed consent was waived by the ethics committee.

Consent for publication

Not applicable.

Data Availability Statement

The datasets generated and/or analyzed during the current study are not publicly available due to institutional data use policies but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was financially supported by two grants: the Hebei Natural Science Foundation (Grant No. H2024112019) and the S&T Program of Xiongan New Area (Grant No. XA202401102001K).

Authors' contributions

Xin Li was responsible for code implementation, data preprocessing, model optimization, manuscript writing, and submission. Yixuan Ding contributed to clinical data collection, assisted with data preprocessing, and participated in manuscript revision and polishing. Bohan Huang participated in data collection, developed inclusion criteria, and contributed to data processing. Yunheng Shen assisted with early-stage code development and model debugging. Hairong Lv supported model tuning and performance optimization. Feng Cao provided clinical supervision, contributed to the study design, and revised the manuscript. Tong Yu polished the manuscript and contributed to language refinement. Fei Li, Xiaolu Fei, and Jia Li served as corresponding authors, supervised the overall project, and provided critical revision of the manuscript.

Acknowledgements

No acknowledgements.

Author details

[1]Department of Biomedical Engineering, Capital Medical University, No. 10, Youanmen Wai Xitoutiao, Beijing 100069, China.

[2]Tsinghua University, No. 1, Tsinghua Garden, Beijing 100084, China.

[3]Xuanwu Hospital of Capital Medical University, No. 45, Changchun Street, Beijing 100053, China.

[4]Xiongan Xuanwu Hospital, No. 1, Xili Road, Xufa District, Xiongan New Area, Hebei 070001, China.

References

[1] Peery, A. F. et al. Burden of gastrointestinal disease in the United States: 2012 update. *Gastroenterology* **143**, 1179–1187.e3 (2012). https://doi.org/10.1053/j.gastro.2012.08.002

[2] Garg, S. K. et al. Incidence, admission rates, and economic burden of adult emergency visits for chronic pancreatitis: Data from the National Emergency Department Sample, 2006 to 2012. *J. Clin. Gastroenterol.* **53**, e328–e333 (2019). https://doi.org/10.1097/MCG.0000000000001096

[3] Banks, P. A. et al. Classification of acute pancreatitis—2012: Revision of the Atlanta classification and definitions by international consensus. *Gut* **62**, 102–111 (2013). https://doi.org/10.1136/gutjnl-2012-302779

[4] Shah, J., Fernandez Y Viesca, M., Jagodzinski, R. and Arvanitakis, M. Infected pancreatic necrosis—current trends in management. *Indian J. Gastroenterol.* **43**, 578–591 (2024). https://doi.org/10.1007/s12664-023-01506-w

[5] de-Madaria, E. and Buxbaum, J. L. Advances in the management of acute pancreatitis. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 691–692 (2023). https://doi.org/10.1038/s41575-023-00808-w

[6] Kumar, A. H. and Griwan, M. S. A comparison of APACHE II, BISAP, Ranson's score and modified CTSI in predicting the severity of acute pancreatitis based on the 2012 revised Atlanta classification. *Gastroenterol. Rep.* **6**, 127–131 (2018). https://doi.org/10.1093/gastro/gox029

[7] Gao, W., Yang, H. X. and Ma, C. E. The value of BISAP score for predicting mortality and severity in acute pancreatitis: A systematic review and meta-analysis. *PLoS One* **10**, e0130412 (2015). https://doi.org/10.1371/journal.pone.0130412

[8] Alberti, P. et al. Evaluation of the modified computed tomography severity index (MCTSI) and computed tomography severity index (CTSI) in predicting severity and clinical outcomes in acute pancreatitis. *J. Dig. Dis.* **22**, 41–48 (2021). https://doi.org/10.1111/1751-2980.12961

[9] Knaus, W. A., Draper, E. A., Wagner, D. P. and Zimmerman, J. E. APACHE II: a severity of disease classification system. *Crit. Care Med.* **13**, 818–829 (1985).

[10] Mao, W. et al. Prediction of infected pancreatic necrosis in acute necrotizing pancreatitis by the modified pancreatitis activity scoring system. *United Eur. Gastroenterol. J.* **11**, 69–78 (2023). https://doi.org/10.1002/ueg2.12353

[11] Wang, L. et al. A simple new scoring system for predicting the mortality of severe acute pancreatitis: a retrospective clinical study. *Medicine (Baltimore)* **99**, e20646 (2020). https://doi.org/10.1097/MD.0000000000020646

[12] Sadr, H., Nazari, M., Khodaverdian, Z. et al. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *Eur. J. Med. Res.* **30**, 418 (2025). https://doi.org/10.1186/s40001-025-02680-7

[13] Thapa, R. et al. Early prediction of severe acute pancreatitis using machine learning. *Pancreatology* **22**, 43–50 (2022). https://doi.org/10.1016/j.pan.2021.10.003

[14] Wiese, M. L. et al. Identification of early predictors for infected necrosis in acute pancreatitis. *BMC Gastroenterol.* **22**, 405 (2022). https://doi.org/10.1186/s12876-022-02490-9

[15] Zhang, H. et al. Tree-based ensemble machine learning models in the prediction of acute respiratory distress syndrome following cardiac surgery: a multicenter cohort study. *J. Transl. Med.* **22**, 772 (2024). https://doi.org/10.1186/s12967-024-05395-1

[16] Muhammad, D. and Bendechache, M. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis. *Comput. Struct. Biotechnol. J.* **24**, 542–560 (2024). https://doi.org/10.1016/j.csbj.2024.08.005

[17] Shakeri, E. et al. Explaining eye diseases detected by machine learning using SHAP: a case study of diabetic retinopathy and choroidal nevus. *SN Comput. Sci.* **4**, 433 (2023). https://doi.org/10.1007/s42979-023-01859-1

[18] Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N. and Narvekar, M. A study of LIME and SHAP model explainers for autonomous disease predictions. *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Mumbai, India, 1–6 (2022). https://doi.org/10.1109/IBSSC56953.2022.10037324

[19] Sahu, B., Abbey, P., Anand, R., Kumar, A., Tomer, S. and Malik, E. Severity assessment of acute pancreatitis using CT severity index and modified CT severity index: correlation with clinical outcomes and severity grading as per the Revised Atlanta classification. *Indian J. Radiol. Imaging* **27**, 152–160 (2017). https://doi.org/10.4103/ijri.IJRI_300_16

[20] Wan, J. et al. Serum D-dimer levels at admission for prediction of outcomes in acute pancreatitis. *BMC Gastroenterol.* **19**, 67 (2019). https://doi.org/10.1186/s12876-019-0989-x

[21] Das, K. et al. Optimized feature-driven dengue diagnosis using explainable machine learning approaches.*Proc. Int. Conf. Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, Rangpur, Bangladesh,**1–6** (2025).https://doi.org/10.1109/QPAIN66474.2025.11171726

[22] Mamun, M., Hussain, M. I., Ali, M. S., Alam Chowdhury, M. S., Chowdhury, S. H. and Hossain, M. M. An explainable ensemble learning framework with feature optimization for accurate maternal health risk prediction. *Proc. Int. Conf. Quantum Photon-*

ics, Artificial Intelligence, and Networking (QPAIN), Rangpur, Bangladesh, **1–6** (2025). https://doi.org/10.1109/QPAIN66474.2025.11172243

[23] Chowdhury, S. H. et al. Hepatitis C detection from blood donor data using hybrid deep feature synthesis and interpretable machine learning. *Proc. 2nd Int. Conf. Next-Generation Computing, IoT and Machine Learning (NCIM)*, Gazipur, Bangladesh, **1–6** (2025). https://doi.org/10.1109/NCIM65934.2025.11160156

[24] Chowdhury, S. H., Mamun, M., Shaikat, T. A., Hussain, M. I., Iqbal, S. and Hossain, M. M. An ensemble approach for artificial neural network-based liver disease identification from optimal features through hybrid modeling integrated with advanced explainable AI. *Medinformatics* **2**, 107–119 (2025). https://doi.org/10.47852/bonviewMEDIN52024744

[25] Mamun, M., Chowdhury, S. H., Hossain, M. M., Khatun, M. R. and Iqbal, S. Explainability-enhanced liver disease diagnosis technique using tree selection and stacking ensemble-based random forest model. *Informatics and Health* **2**, 17–40 (2025). https://doi.org/10.1016/j.infoh.2025.01.001

[26] Párniczky, A. *et al.* International Association of Pancreatology Revised Guidelines on Acute Pancreatitis 2025: Supported and Endorsed by the American Pancreatic Association, European Pancreatic Club, Indian Pancreas Club, and Japan Pancreas Society. *Pancreatology* **25**(6), 770–814 (2025). https://doi.org/10.1016/j.pan.2025.04.020