Development and validation of an explainable machine learning model for mortality prediction among patients with infected pancreatic necrosis

Caihong Ning,[a,b,c,d,e,g] Hui Ouyang,[a,b,g] Jie Xiao,[f] Di Wu,[f] Zefang Sun,[a,b,c,d] Baiqi Liu,[a,b,c,d] Dingcheng Shen,[a,b,c,d] Xiaoyue Hong,[a,b,c,d] Chiayan Lin,[a,b,c,d] Jiarong Li,[a,b,c,d] Lu Chen,[a,b,c,d] Shuai Zhu,[a,b,c,d] Xinying Li,[a,b] Fada Xia,[a,b,**] and Gengwen Huang[a,b,c,d,e,*]

[a]Department of General Surgery, Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China
[b]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China
[c]Division of Pancreatic Surgery, Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China
[d]Division of Hernia and Abdominal Wall Surgery, Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China
[e]FuRong Laboratory, Changsha, Hunan Province 410078, China
[f]Department of Emergency, Third Xiangya Hospital, Central South University, Changsha, Hunan Province 410008, China

eClinicalMedicine
2025;80: 103074
Published Online xxx
https://doi.org/10.1016/j.eclinm.2025.103074

Summary

Background Infected pancreatic necrosis (IPN) represents a severe complication of acute pancreatitis, commonly linked with mortality rates ranging from 15% to 35%. However, the present mortality prediction tools for IPN are limited and lack sufficient sensitivity and specificity. This study aims to develop and validate an explainable machine learning (ML) model for death prediction among patients with IPN.

Methods We performed a prospective cohort study of 344 patients with IPN consecutively enrolled from a large Chinese tertiary hospital from January 2011 to January 2023. Ten ML models were developed to predict 90-day mortality in these patients. A benchmarking test, involving nested resampling, automatic hyperparameter tuning and random search techniques, was conducted to select the ML model. Sequential forward selection method was employed to select the optimal feature subset from 31 candidate subsets to simplify the model and maximize predictive performance. The final model was internally validated with the 1000 bootstrap method and externally validated using an independent cohort of 132 patients with IPN retrospectively collected from another Chinese tertiary hospital from January 2018 to January 2023. The SHapley Additive exPlanations (SHAP) method was employed to interpret the model in terms of features importance and features effect. The final model constructed with optimal feature subset was deployed as an interactive web-based Shiny app.

Findings Random survival forest (RSF) model showed the best predictive performance than other 9 ML models (internal validation, C-index = 0.863 [95% CI: 0.854–0.875]; external validation, C-index = 0.857 [95% CI: 0.850–0.865]). Multiple organ failure, Acute Physiology and Chronic Health Examination II (APACHE II) score ≥20, duration of organ failure ≥21 days, bloodstream infection, time from onset to first intervention <30 days, Bedside Index of Severity in Acute Pancreatitis score ≥3, critical acute pancreatitis, age ≥ 50 years, and hemorrhage were 9 most important features associated with mortality. Furthermore, SHAP algorithm revealed insightful nonlinear interactive associations between important predictors and mortality, identifying 9 features pairs with high interaction SHAP value and clinical significance. Two interactive web-based Shiny apps were developed to enhance clinical practicability: https://rsfmodels.shinyapps.io/IPN_app/ for cases where the APACHE II score was available and https://rsfmodels.shinyapps.io/IPNeasy/ for cases where it was not.

Interpretation An explainable ML model for death prediction among IPN patients was feasible and effective, suggesting its superior potential in guiding clinical management and improving patient outcomes. Two publicly accessible web tools generated for the optimized model facilitated its utility in clinical settings.

*Corresponding author. Department of General Surgery, Xiangya Hospital, Central South University, 87 Xiangya Rd, Changsha, Hunan Province 410008, China.
**Corresponding author. Department of General Surgery, Xiangya Hospital, Central South University, 87 Xiangya Rd, Changsha, Hunan Province 410008, China.
E-mail addresses: huanggengwen@csu.edu.cn (G. Huang), xiafada@xiangya.com.cn (F. Xia).
[g]Contributed equally to this work.

Funding The Natural Science Foundation of Hunan Province (2023JJ30885), Postdoctoral Fellowship Program of CPSF (GZB20230872), The Youth Science Foundation of Xiangya Hospital (2023Q13), The Project Program of National Clinical Research Center for Geriatric Disorders of Xiangya Hospital (2021LNJJ19).

Keywords: Infected pancreatic necrosis; Machine learning; Prediction model; SHapley additive exPlanations; Shiny app

Research in context

Evidence before this study

Before the start of our study, we did an extensive literature search using PubMed, Scopus, and Medline, focusing on studies published up to 20 November 2024. We used the search terms "pancreatitis", "machine learning" and "mortality". We excluded nine cohort studies that employed machine learning for mortality prediction in patients with acute pancreatitis. Ultimately, only one study utilizing machine learning for mortality prediction in patients with infected pancreatic necrosis was identified. This study analyzed 223 patients who underwent surgery for infected pancreatic necrosis at West China Hospital of Sichuan University. The results indicated that preoperative modified Marshall score, time of surgery, duration of organ failure, and onset of renal failure were important predictive factors for postoperative mortality in patients undergoing delayed surgery (≥4 weeks). However, the study was limited by its retrospective, single-center design and lacked external validation. Additionally, the machine learning model was lack of explanation and deployment.

Added value of this study

To our knowledge, this is the first and largest study to compare 10 machine learning models and develop an explainable machine learning model with optimal predictive performance for mortality among patients with infected pancreatic necrosis. Additionally, this study is the first to offer two publicly accessible web tools to facilitate the clinical utility of the machine learning-based mortality prediction model for infected pancreatic necrosis.

Implications of all the available evidence

Based on the results of this study, we expect that clinicians can identify patients with infected pancreatic necrosis at high risk of death early by using two interactive web-based Shiny apps for random survival forest model. Furthermore, multiple organ failure, Acute Physiology and Chronic Health Examination II score ≥20, duration of organ failure ≥21 days, bloodstream infection, time from onset to first intervention <30 days, Bedside Index of Severity in Acute Pancreatitis score ≥3, critical acute pancreatitis, age ≥ 50 years, and hemorrhage were 9 most important features associated with increased mortality. Therefore, in clinical practice, persistent organ failure should be reversed within 21 days, and surgical intervention should, whenever possible, be delayed until at least 30 days from the onset. The step-down approach should be avoided whenever possible in cases of critical acute pancreatitis, particularly in patients with multiple organ failure and a duration of organ failure ≥21 days. However, future studies with larger cohorts and more diverse external validation, incorporating precision medicine techniques and advanced artificial intelligence, would be needed to enhance the applicability and generalizability of these findings.

Introduction

Acute pancreatitis (AP) is one of the most common gastrointestinal disorders requiring acute hospital admission.[1,2] The global estimates indicate an incidence of 34 cases and a mortality rate of 2 cases per 100,000 person-years.[3] The majority of AP cases exhibit mild symptoms, characterized by a self-limiting course.[1,4] However, approximately 20% of patients progress to moderate or severe acute pancreatitis, involving pancreatic or peripancreatic necrosis and/or organ failure.[1,4] Among these cases, 67% exhibit sterile pancreatic necrosis, while 33% present with infected pancreatic necrosis (IPN), a significant contributor to mortality, with mortality rates up to 15%–35%.[4] In recent years, despite advances in critical care and minimally invasive techniques, the mortality of IPN has remained around 15–20% or higher, even in specialized centers. Therefore, timely and accurate identification of high-risk patients is crucial for guiding clinical management, so as to enhance the prognosis of patients with IPN.

Several biochemical markers and scoring systems have been developed to predict the severity and mortality of AP, such as c-reactive protein, blood urea nitrogen, Acute Physiology and Chronic Health Examination II (APACHE II), Bedside Index of Severity in Acute Pancreatitis (BISAP), Harmless Acute Pancreatitis Score (HAPS), and CT Severity Index (CTSI), modified CT Severity Index (MCTSI).[2,5–8] Recently, traditional models in several studies have identified certain mortality predictors for IPN.[9–11] However, the accuracy and specificity of these scoring systems and mortality predictors remain unsatisfactory. Machine learning (ML), as a

branch of artificial intelligence, can detect complex, nonlinear relationships between various features and disease outcomes, and has been widely applied in the field of disease diagnosis, complication monitoring, and prognosis prediction to assist physicians in decision-making.[12–18] Several studies have employed ML to establish predictive tools for AP, mainly focused on severity, complication, and mortality.[19–23] However, most studies were retrospective and single-center, featuring relatively small patient populations and lacking external validation and model interpretability. In our previous research, 7 most important predictors has been identified to have nonlinear relationship with mortality of IPN through the random survival forest (RSF) algorithm.[24] However, the study lacked the development of an explainable model for death prediction, and external validation was not performed.

This study aimed to develop and validate an explainable ML model for accurately predicting the individual-level risk of death in IPN cases, with SHapley Additive exPlanations (SHAP) method demonstrating the effect of important features and potential interactive effects between features.[25] Additionally, an interactive web-based Shiny app for the model was designed to enhance its applicability in clinical settings.

Methods

Study design and population

Two different datasets were used to develop and validate the explainable ML model for death prediction among IPN patients, respectively (Fig. 1). The derivation cohort was a prospective cohort of 344 patients with IPN consecutively enrolled from Xiangya hospital, Central South University, China between January 2011 and January 2023, while the external validation cohort was a retrospective cohort of 132 consecutive patients with IPN admitted to the Third Xiangya Hospital, Central South University, China between January 2018 and January 2023. The exclusion criteria for the derivation cohort included patients with a history of chronic pancreatitis (n = 4), patients with chronic organ dysfunction (n = 6), patients during pregnancy (n = 2), and patients with incomplete data (n = 8). The exclusion criteria for the validation cohort included patients with a history of chronic pancreatitis (n = 3), patients with chronic organ dysfunction (n = 5), patients during pregnancy (n = 3), and patients with incomplete data (n = 14). Patients with missing values were considered to have incomplete data and were excluded from the study. The study was approved by the Ethics Committees of Xiangya Hospital (No.201012067) and the Third Xiangya Hospital (No.21019) and registered www.researchregistry.com (https://www.researchregistry.com/register-now#home/registrationdetails/64b8b18bc0679a0027c1e25b/). Routine written informed consent was obtained from all participants or their legal representatives for the collection and publication of data. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement and the reporting guideline of Strengthening The Reporting Of Cohort Studies in Surgery (STROCSS).[26,27]

Data preprocessing and candidate variables

Data were extracted from two IPN cohorts, collected both prospectively and retrospectively by trained professionals, and sourced from the electronic medical record systems of two hospitals. Thirty-one candidate variables were collected based on data availability and clinical knowledge that spanned demographic, clinical, and treatment and complication-related attributes: 9 demographic and baseline variables (age, gender, comorbidity, smoking or drinking, etiology, severity classification, intensive care unit (ICU) stay, the number and duration of organ failure), 4 scoring systems variables (APACHEII score, BISAP score, CTSI score, MCTSI score), 2 therapeutic variables (time from onset to first intervention, step-up or step-down surgical approach), 4 complication variables (gastrointestinal fistula, hemorrhage, gastrointestinal fistula or hemorrhage, pancreatic fistula), 12 infection-associated variables (pancreatic polymicrobial infection, pancreatic *Klebsiella pneumoniae* infection, pancreatic *Acinetobacter baumannii* infection, pancreatic *Enterococcus faecium* infection, pancreatic *Escherichia coli* infection, pancreatic fungal infection, pancreatic Carbapenem-resistant Enterobacter (CRE) infection, pancreatic multidrug-resistant organisms (MDRO) infection, bloodstream infection, bloodstream MDRO infection, bloodstream CRE infection, candidemia). Most of the candidate variables were collected at the baseline of IPN diagnosis except for step-up or step-down surgical approach. The primary outcome was death occurring within 90-day from the onset of the disease. The interval, measured in days, from the initial diagnosis to the recorded date of death was defined as overall survival. The definition of other variables were detailed in the Supplemental methods.

Model and feature selection

The workflow of ML was showed in Fig. 1. A total of 10 ML models were developed for the death prediction of patients with IPN: coxph, glmnet, rpart, RSF, gbm, svm, xgboost, deepsurv, deephit, coxtime (Fig. 2). All the ten ML algorithm employed in our research was specifically based on survival analysis (Supplemental Table S1). The svm model referred specifically to survivalsvm. To obtain an unbiased and objective evaluation of multiple ML models under consistent conditions, a benchmarking test involving nested resampling, automatic hyperparameter tuning and random search techniques was designed. The selection of the optimal model from 10 models was based on a

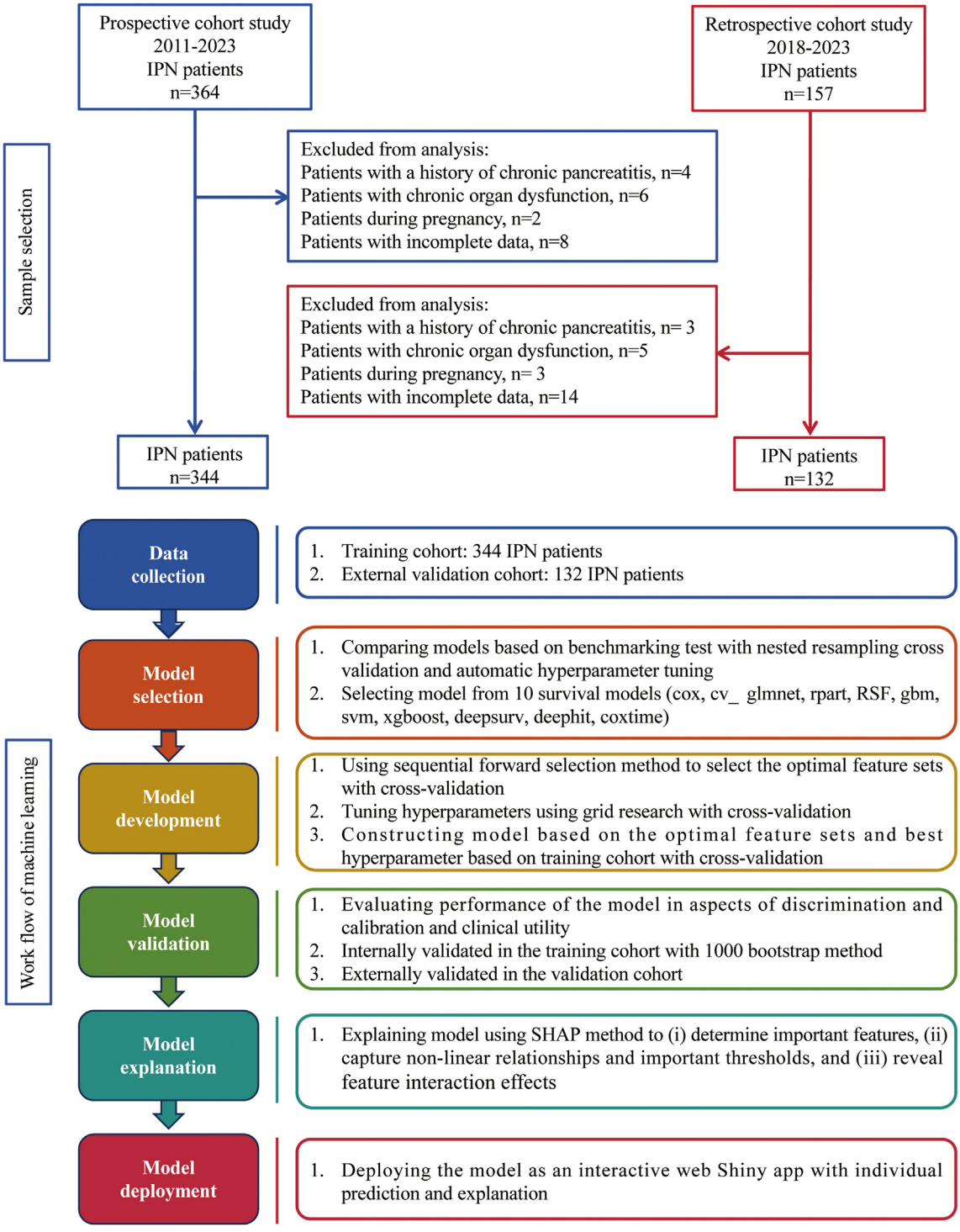

Fig. 1: Overview of the study workflow. IPN, infected pancreatic necrosis; RSF, random survival forest; SHAP, SHapley Additive exPlanations.

comprehensive evaluation of the concordance index (C-index) and Brier score. To simplify the model for enhanced applicability, the sequential forward selection method was employed to select the optimal feature subset based on the selected model above. Notably, the optimal feature subset was selected by striking a balance between

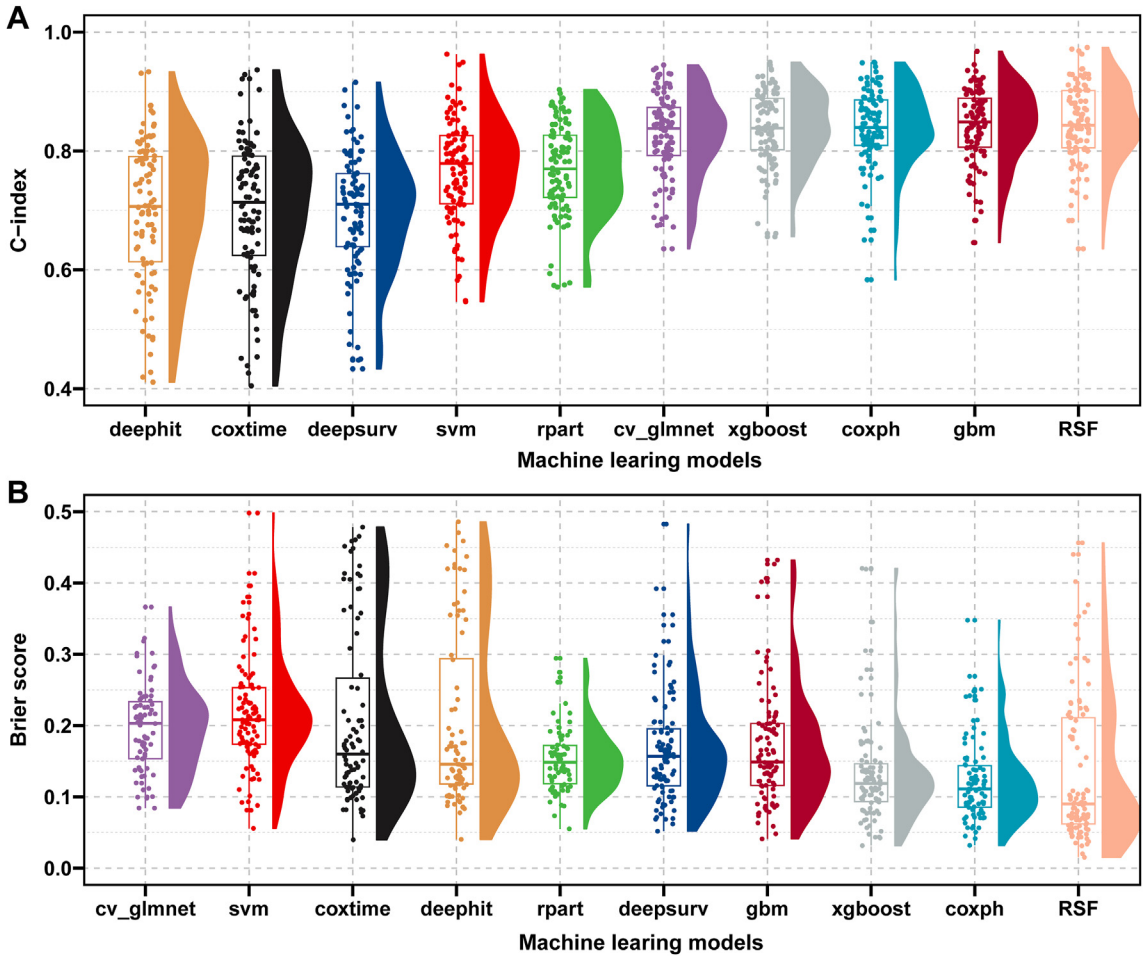

Fig. 2: Model selection from 10 machine learning models based on C-index (A) and Brier score (B). RSF, random survival forest.

achieving a high C-index and maintaining a manageable number of features (Supplemental methods, Figs. S1–S3, and Tables S1 and S2).

Model construction and validation

The final model was developed based on the optimal model and feature subset above. The performance of the final model was assessed in terms of discrimination, calibration, and clinical utility. Discrimination ability was measured using C-index and time-dependent area under curve (AUC). Calibration capability was assessed through calibration curves and the integrated Brier score. Of note, the C-index measures a model's ability to discriminate between subjects who experienced the event of interest vs. those who did not, while the Brier score assesses the calibration and goodness-of-fit of the predicted survival probabilities from a model.[28] Moreover, clinical utility was evaluated using decision curve analysis (DCA). The final model was validated internally and externally. The internal validation was performed on the derivation cohort using the 1000 bootstrap method, while the external validation was conducted on an independent test cohort. Meanwhile, we have also developed the final model with all features and assessed its performance internally and externally. More details of model construction and validation were provided in Supplemental methods.

Model explanation and deployment

SHAP, introduced by Lundberg and Lee, offers a novel approach to elucidate predictions generated by various black-box machine learning models.[29] In the study, the SHAP summary and dependence plots were employed to identify key predictors and investigate their relationships with the outcome. Additionally, the SHAP interactive plot was used to identify potential interaction effects between two features. To facilitate the accessibility and usability of the model, the final model was deployed as an interactive web-based Shiny app that enables individualized survival prediction, personalized

interpretation, and explanation of the model. More information was described in Supplemental methods.

Sensitivity and statistical analysis

Several procedures of sensitivity analysis were employed to assess the robustness of results in the training cohort: (1) explaining the final model (developed from the training cohort) in the validation cohort; (2) assessing feature importance with the permutation method; (3) performing interaction effect analysis with the traditional Cox model.

Summary statistics were presented as total frequencies and percentages for categorical variables, and reported as median with interquartile range (IQR) or as means and standard deviations (SD) for continuous variables, as appropriate. Differences in data distribution between datasets for both categorical and continuous variables were assessed by the $\chi 2$ test and Mann–Whitney U test, respectively, with a 2-sided P value < 0.05 considered statistically significant. R version 4.3.1 was used to perform all statistical analyses and create all figures. A list of the R statistical packages utilized for the analyses in R version 4.3.1 is detailed in the supplemental file (Supplemental Table S3).

Role of the funding source

The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Results

Patients characteristics

Clinical characteristics of the study population were presented in Table 1. There were 344 patients in the derivation cohort and 132 patients in the external validation cohort, in which death occurred in 83 (24.1%) and 42 (31.8%) patients, respectively. In the derivation cohort, 73.0% were males and 27.0% were females. The median age of the cohort was 48 (37–55) years. The derivation and validation cohorts were similar in terms of gender, age, comorbidity, smoking or drinking, etiology, IPN patients diagnosed on admission, ICU stay, time from onset to first intervention, surgery approach, BISAP score, CTSI score, MCTSI score, gastrointestinal fistula, hemorrhage, gastrointestinal fistula or hemorrhage, pancreatic fistula, pancreatic and bloodstream infection, and 90-day mortality. However, the validation cohort had a significantly higher rate of critical acute pancreatitis and multiple organ failure (MOF), longer median duration of organ failure, and higher APACHE II score (P < 0.05).

Model selection and construction

Ten ML models were developed for death prediction of patients with IPN based on 31 candidate features in the derivation cohort. After a benchmarking test, the RSF algorithm achieved the best predictive performances with the highest mean (SD) C-index of 0.865 (0.066) and the lowest mean (SD) Brier score of 0.147 (0.121), outperforming other models (C-index ranging from 0.698 to 0.863, Brier score ranging from 0.151 to 0.353) (Fig. 2 and Supplemental Table S4). Consequently, the RSF model was used to construct the final models. Next, the sequential forward selection method was performed to identify optimal feature subsets that maximized the performance in the RSF model (Supplemental Fig. S2). As a result, the final model achieving the optimal C-index was constructed by 10 features, including age, smoking or drinking, APACHE II score, number of organ failures, duration of organ failure, bloodstream infection, pancreatic CRE infection, time from onset to first intervention, surgery approach and hemorrhage (Supplemental Fig. S3 and Table S5). Additionally, the RSF model with all features was developed for model explanation. Both RSF models, whether utilizing the optimal features or all features, underwent hyperparameter optimization (Supplemental Fig. S4).

Model validation

Fig. 3 showed the excellent performance of the final RSF model with the optimal features both in the derivation and external validation cohort regarding discrimination, accuracy, and clinical applicability. In the derivation cohort, discrimination, evaluated through time-dependent AUC analyses, consistently demonstrated high values, reflecting sustained discriminative ability over time. The C-index was 0.863 (95% CI: 0.854–0.875) with 1000 bootstrap resampling (Fig. 3A), indicating good discriminatory ability. Simultaneously, the calibration plot was used to assess the predicted accuracy of 30-day, 60-day, and 90-day overall survival, revealing a noteworthy correspondence with the ideal curve (Fig. 3B). The integrated Brier score was 0.153 (95% CI:0.143–0.163) with 1000 bootstrap resampling, further endorsing the model's high reliability. Moreover, DCA affirmed the RSF model's commendable clinical applicability as a tool for initiating medical intervention (Fig. 3C). Importantly, in the external validation cohort, the time-dependent AUC curves, calibration plots, and DCA curves to evaluate the RSF model were presented in Fig. 3D–F, with C-index for 0.857 (95% CI: 0.850–0.865) and Brier score for 0.084 (95% CI: 0.076–0.092). Notably, the performance of the RSF model with all features showed similar results with the RSF model with the optimal features (Supplemental Fig. S5).

Model explanation and deployment

The SHAP method was utilized to explain the model's explainability by generating an importance ranking of candidate features on individual predictions. As shown in Fig. 4, the top 9 most important prognostic features

Characteristics	Levels	Derivation cohort (n = 344)	Validation cohort (n = 132)	*P* value
Dead	No	261 (75.9%)	90 (68.2%)	0.112
	Yes	83 (24.1%)	42 (31.8%)	
Gender	Female	93 (27%)	39 (29.5%)	0.665
	Male	251 (73%)	93 (70.5%)	
Age	Median (IQR)	48 (37–55)	48 (42–52.5)	0.862
Complications	No	208 (60.5%)	67 (50.8%)	0.069
	Yes	136 (39.5%)	65 (49.2%)	
Smoking or drinking	No	191 (55.5%)	79 (59.8%)	0.454
	Yes	153 (44.5%)	53 (40.2%)	
Etiology	Biliary	127 (36.9%)	43 (32.6%)	0.739
	hypertriglyceridemia	150 (43.6%)	62 (47%)	
	Alcoholic	20 (5.8%)	10 (7.6%)	
	Other	47 (13.7%)	17 (12.9%)	
Infected pancreatic necrosis diagnosed on admission	Yes	137 (39.8)	59 (44.7)	0.334
	No	207 (60.2)	73 (55.3)	
ICU stay	No	90 (26.2%)	33 (25%)	0.887
	Yes	254 (73.8%)	99 (75%)	
Severity classification	Severe acute Pancreatitis	165 (48%)	49 (37.1%)	0.043
	Critical acute Pancreatitis	179 (52%)	83 (62.9%)	
Number of organ failures	No	155 (45.1%)	50 (37.9%)	0.022
	Single organ failure	71 (20.6%)	19 (14.4%)	
	Multiple organ Failure	118 (34.3%)	63 (47.7%)	
Duration of organ failure	Median (IQR)	2 (0–18)	10 (0–23)	0.013
Time from onset to first intervention	Median (IQR)	21.50 (14–31.5)	22 (18.5–30)	0.120
Surgery approach	Step-up approach	270 (78.5%)	105 (79.5%)	0.899
	Step-down approach	74 (21.5%)	27 (20.5%)	
APACHE II score	Median (IQR)	8 (5–15)	9 (7–14)	0.025
BISAP score	Median (IQR)	2 (2–3)	2 (2–3)	0.395
MCTSI score	Median (IQR)	10 (8–10)	10 (8–10)	0.133
CTSI score	Median (IQR)	8 (6–10)	8 (6–10)	0.064
Gastrointestinal fistula	No	291 (84.6%)	107 (81.1%)	0.427
	Yes	53 (15.4%)	25 (18.9%)	
Hemorrhage	No	269 (78.2%)	94 (71.2%)	0.138
	Yes	75 (21.8%)	38 (28.8%)	
Gastrointestinal fistula or Hemorrhage	No	236 (68.6%)	87 (65.9%)	0.650
	Yes	108 (31.4%)	45 (34.1%)	
Pancreatic fistula	No	192 (55.8%)	67 (50.8%)	0.374
	Yes	152 (44.2%)	65 (49.2%)	
Peripancreatic polymicrobial infection	No	119 (34.6%)	43 (32.6%)	0.758
	Yes	225 (65.4%)	89 (67.4%)	
Pancreatic *Klebsiella pneumoniae*	No	214 (62.2%)	76 (57.6%)	0.411
	Yes	130 (37.8%)	56 (42.4%)	
Pancreatic *Acinetobacter baumannii*	No	249 (72.4%)	92 (69.7%)	0.639
	Yes	95 (27.6%)	40 (30.3%)	
Pancreatic *Enterococcus faecium*	No	241 (70.1%)	89 (67.4%)	0.655
	Yes	103 (29.9%)	43 (32.6%)	
Pancreatic *Escherichia coli*	No	256 (74.4%)	93 (70.5%)	0.447
	Yes	88 (25.6%)	39 (29.5%)	
Pancreatic fungal infection	No	253 (73.5%)	94 (71.2%)	0.691
	Yes	91 (26.5%)	38 (28.8%)	
Pancreatic CRE infection	No	237 (68.9%)	86 (65.2%)	0.501
	Yes	107 (31.1%)	46 (34.8%)	
Pancreatic MDRO infection	No	153 (44.5%)	53 (40.2%)	0.454
	Yes	191 (55.5%)	79 (59.8%)	
Bloodstream infection	No	233 (67.7%)	84 (63.6%)	0.459
	Yes	111 (32.3%)	48 (36.4%)	

(Table 1 continues on next page)

Characteristics	Levels	Derivation cohort (n = 344)	Validation cohort (n = 132)	*P* value
(Continued from previous page)				
Bloodstream MDRO infection	No	275 (79.9%)	97 (73.5%)	0.161
	Yes	69 (20.1%)	35 (26.5%)	
Bloodstream CRE infection	No	309 (89.8%)	113 (85.6%)	0.255
	Yes	35 (10.2%)	19 (14.4%)	
Candidemia	No	328 (95.3%)	126 (95.5%)	1
	Yes	16 (4.7%)	6 (4.5%)	

APACHE II score, Acute Physiology and Chronic Health Examination II score; BISAP score, Bedside Index of Severity in Acute Pancreatitis score; MDRO, Multidrug-resistant organisms; CRE, Carbapenem-resistant Enterobacter; ICU, Intensive care unit; MCTSI, Modified CT Severity Index; CTSI, CT Severity Index; IQR, Interquartile range.

Table 1: **Comparison of baseline characteristics of patients with infected pancreatic necrosis in the derivation cohort and validation cohort.**

contributed to the high likelihood of death were number of organ failures, APACHE II score, duration of organ failure, bloodstream infection, time from onset to first intervention, BISAP score, severity classification, age and hemorrhage.

Additionally, SHAP dependence plots were used to interpret the relationships of critical predictors with the outcome. From the main and total effect plots (Fig. 5 and Supplemental Fig. S6), MOF, positive bloodstream infection, BISAP score ≥3, critical acute pancreatitis, and hemorrhage were unsurprisingly associated with increased risk of death. Notably, we observed a positive linear correlation between age and death risk, potential nonlinear relationships and important thresholds between APACHE II score, duration of organ failure, time from onset to first intervention, and death risk. Specifically, the death risk was not increased when the APACHE II score was less than 10, slightly increased when between 10 and 20, and significantly increased when greater than 20. Additionally, the mortality risk was much higher when the duration of organ failure was longer than approximately 21 days and much lower when time from onset to first intervention was longer than about 30 days.

From the heat map of interaction SHAP value (Fig. 6), there were strong interaction effects between number of organ failures, duration of organ failure, time from onset to first intervention, bloodstream infection, BISAP score, APACHEII score, and severity classification. We then selected the top 9 feature pairs

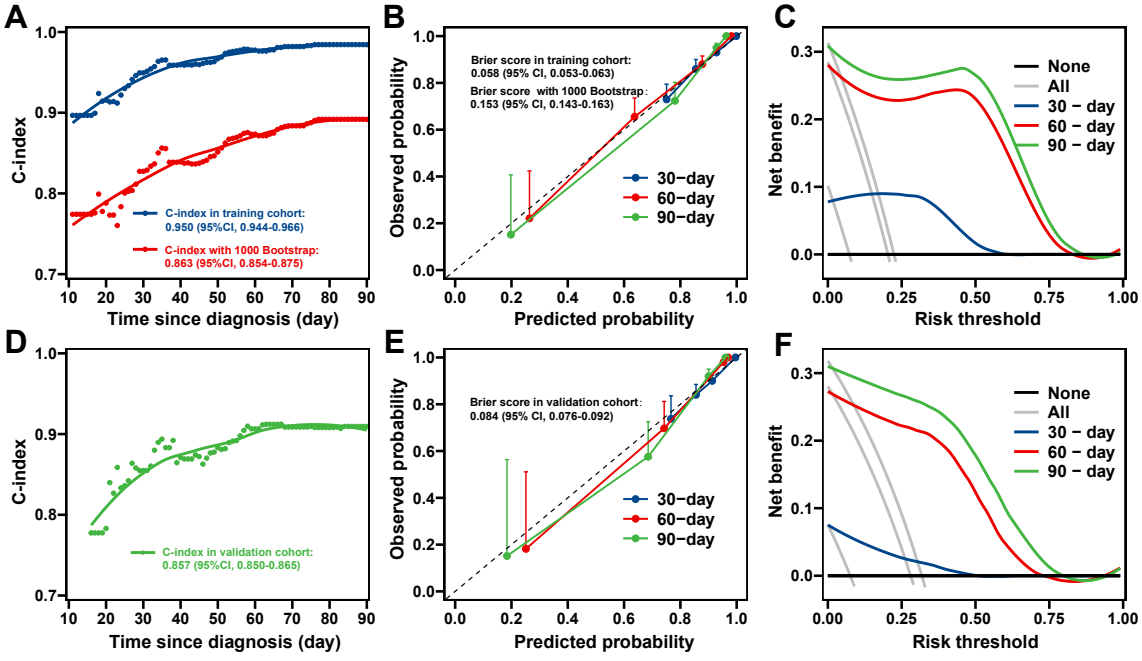

Fig. 3: Performance of the random survival forest model with the best feature set in the training (A, B, C) and external validation (D, E, F) cohort. The model performance was comprehensively visualized with time-dependent area under curve (A, D), calibration plot (B, E), and decision curve analysis plot (C, F).

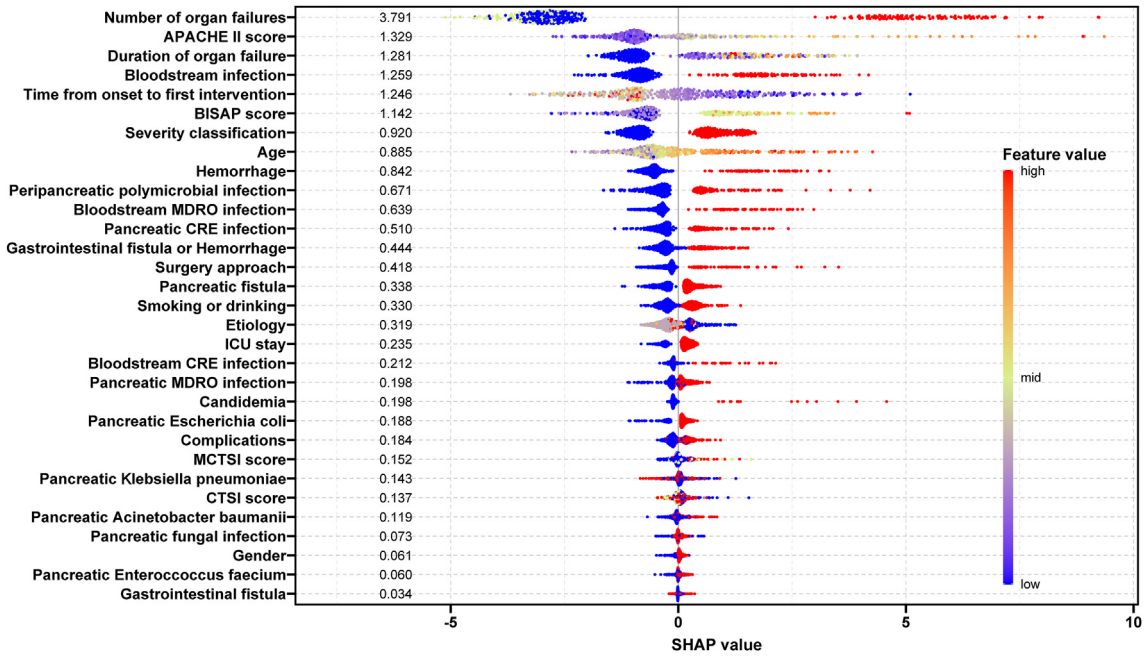

Fig. 4: SHAP summary plot of random survival forest model with all feature in the training cohort. Each dot represented the value of an individual patient data point in the training cohort, with feature's value ranging from low (in blue) to high (in red). The distance of each dot from the center of the x-axis represents the magnitude of impact (total SHAP value) on the model's output, with SHAP value above zero indicating contribution to death (increased death risk), and SHAP value below zero suggesting contribution to survival (reduced death risk). Features were ranked on the y-axis from the highest to the lowest average contribution (average absolute SHAP value) in terms of feature importance. SHAP, SHapley Additive exPlanations; APACHE II score, Acute Physiology and Chronic Health Examination II score; BISAP score, Bedside Index of Severity in Acute Pancreatitis score; MDRO, Multidrug-resistant organisms; CRE, Carbapenem-resistant Enterobacter; ICU, Intensive care unit; MCTSI, Modified CT Severity Index; CTSI, CT Severity Index.

with both high interaction SHAP value and clinical significance (Figs. 6 and 7, and Supplemental Fig. S7). For instance, as shown in Fig. 7A, the negative interaction (SHAP interaction values < 0, indicating better prognosis) was between age < 50 years and MOF, however, this pattern was reversed between age ≥ 50 years and MOF. The positive interaction effect suggested that patients with age ≥ 50 years and MOF would fare worse than expected from the additive prognostic effect of the two variables. Interestingly, the negative interaction effect between age < 50 years and MOF, and positive effect between age ≥ 50 years and MOF could be observed in dependence plots of total effect (Supplemental Fig. S6). Moreover, the addictive positive interaction effect between age ≥ 50 years and MOF was further confirmed by interaction metrics in traditional Cox model (Supplemental Fig. S8 and Tables S6 and S7). More detailed explanation has been provided in Supplemental methods.

Finally, we implemented the final RSF model into an interactive web-based Shiny app that provided survival prediction and explanation for individuals. Additionally, it also provided a global explanation of the model. The web application was made accessible at https://rsfmodels.shinyapps.io/IPN_app/. Given that collecting all the required variables for the APACHE II score may be challenging and may limit the clinical applicability of the models, we developed an alternative online model without the APACHE II score system (https://rsfmodels.shinyapps.io/IPNeasy/). This model achieved a C-index of 0.855 (95% CI: 0.845–0.865) based on 1000 bootstrap samples and 0.851 (95% CI: 0.840–0.861) in the validation cohort (Supplemental Table S8), both of which were slightly lower than the corresponding C-index values of the model incorporating the APACHE II score system.

Sensitivity analysis

The RSF model explanation in the validation cohort indicated similar results in terms of feature importance and feature effect (Supplemental Figs. S9–14). Additionally, feature importance achieved a similar ranking trend across the SHAP method and permutation method both in the training cohort (Supplemental Fig. S15) and validation cohort (Supplemental Fig. S16). Moreover, a similar interaction pattern between age and number of organ failure was seen in the validation cohort (Supplemental Fig. S17 and Tables S9 and S10).

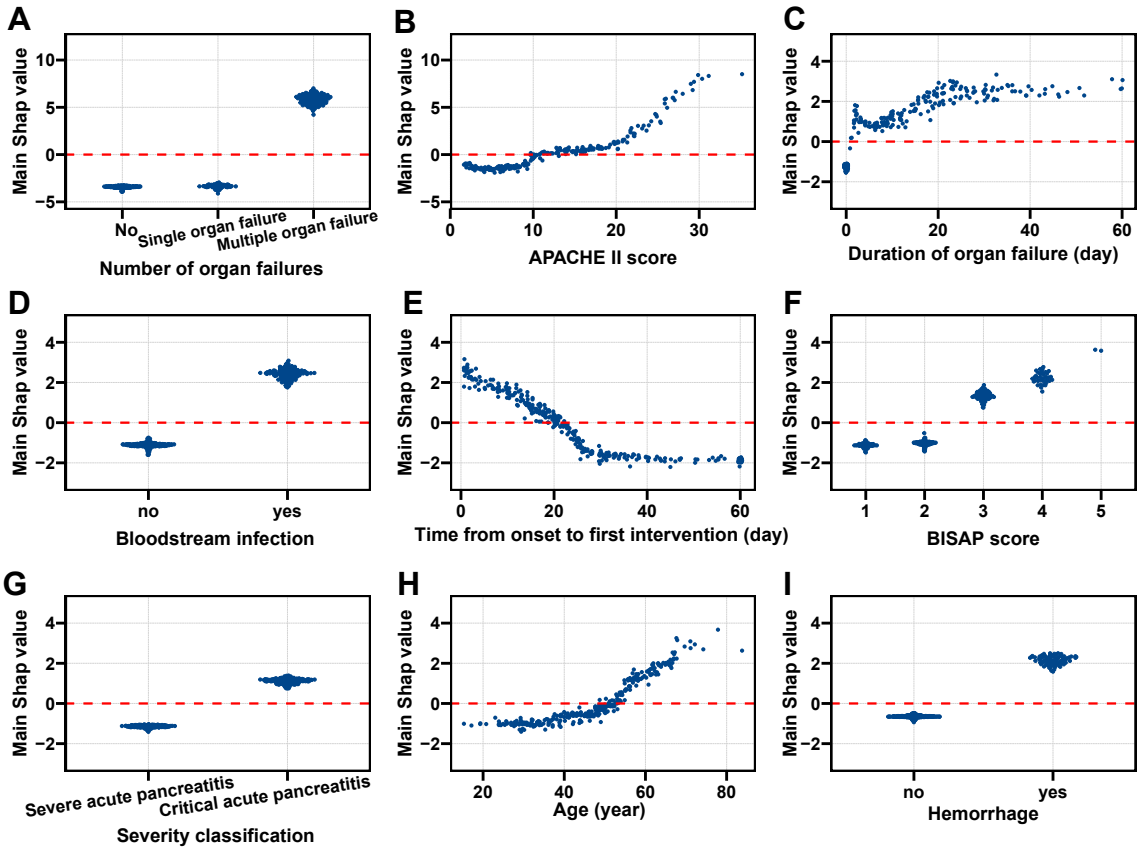

Fig. 5: SHAP dependence plot of top nine important features based on main SHAP value in the training cohort. Plots showed the main effect of a feature on death, including number of organ failures (A), APACHE II score (B), duration of organ failure (C), bloodstream infection (D), time from onset to first intervention (E), BISAP score (F), severity classification (G), age (H), hemorrhage (I). SHAP, SHapley Additive exPlanations; APACHE II score, Acute Physiology and Chronic Health Evaluation II score; BISAP score, Bedside Index for Severity in Acute Pancreatitis score.

Discussion

This is the first and largest study, to our knowledge, to investigate and compare 10 ML models for death prediction analyses in IPN cohorts. The RSF model achieved the best predictive performance both in the derivation and external validation cohort regarding discrimination, accuracy, and clinical applicability. Meanwhile, the SHAP algorithm identified the top 9 contributing factors associated with increased mortality likelihood, revealing insightful nonlinear interactive associations between predictors and death. Furthermore, two publicly accessible web tools were constructed for the optimized model, enhancing its utility in clinical settings.

AP exhibited a diverse clinical course influenced by individual characteristics, ranging from a mild, self-limiting disease to a severe, life-threatening illness with IPN and/or persistent (≥48 h) organ failure.[30] Hence, it is crucial to identify mortality predictors and construct a death prediction model for IPN to guide clinical management and enhance prognosis. Previous studies have highlighted laboratory variables like c-reactive protein, along with scoring systems such as APACHE II, Ranson, and SOFA, for assessing the severity and prognosis of AP.[2,5–8] However, these parameters have limitations, requiring an extensive set of mandatory variables or exhibiting low sensitivity and specificity for mortality prediction, thus limiting their clinical utility. Moreover, existing scoring systems were typically effective at admission but lacked utility in the late phase of the disease due to the varied clinical course of AP.

Several studies have developed ML death prediction models for AP. Ding et al. developed an artificial neural network prediction model for in-hospital mortality based on age and 11 laboratory biochemical variables (e.g., alanine aminotransferase, total bilirubin, creatine kinase isoenzyme, prothrombin time, white blood cells, amylase, total calcium, creatinine, hematocrit, lactate, and lipase) within 24 h after admission of AP patients in the Medical Information Mart for Intensive Care III database, achieving an AUC of 0.769, notably higher

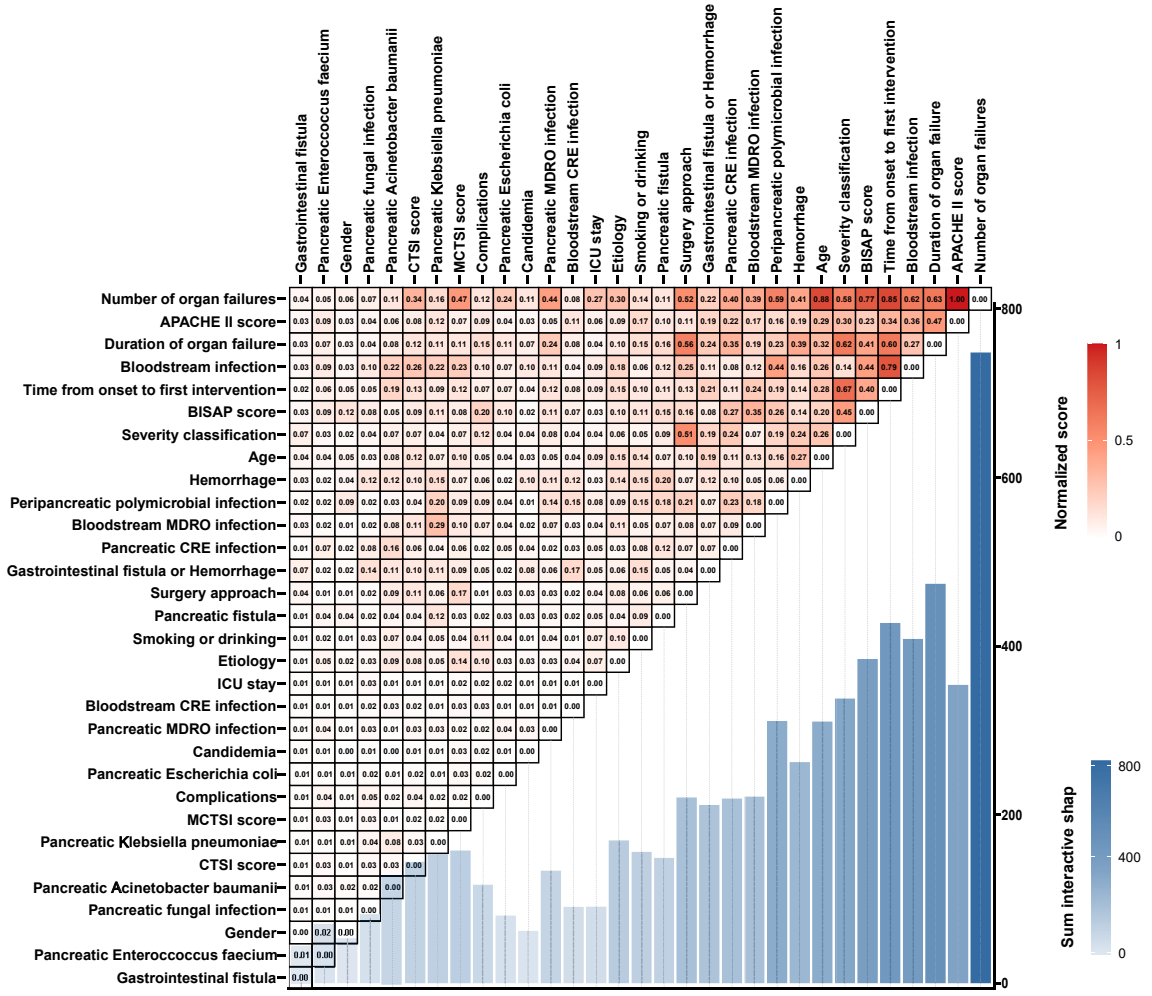

Fig. 6: Heatmap of interaction effect between feature pairs in the training cohort. The value indicated the normalized absolute interaction SHAP value, with higher score (darker red color) indicating higher interaction effect between the feature pair. The bar plot suggested the sum absolute interaction SHAP value of a feature with other features, with darker blue color indicating higher sum interaction effect. SHAP, SHapley Additive exPlanations; APACHE II score, Acute Physiology and Chronic Health Examination II score; BISAP score, Bedside Index of Severity in Acute Pancreatitis score; MDRO, Multidrug-resistant organisms; CRE, Carbapenem-resistant Enterobacter; ICU, Intensive care unit; MCTSI, Modified CT Severity Index; CTSI, CT Severity Index.

than the Ranson score.[22] However, it was a retrospective public database study with relatively small samples, limited clinical and laboratory characteristics, and inevitable missing part of data. Moreover, as all variables were collected during the early stage of AP, applying the model in later stage of the disease might pose limitations, which was similar with the limitations of ANN model for mortality prediction in AP developed by Mofidi et al. [31] Lan et al. focused on the impact of surgical timing on mortality, analyzing 223 IPN patients who underwent surgery at West China Hospital.[23] They identified the key factors associated with surgical timing (<4 or ≥4 weeks) and postoperative mortality for IPN and predicted the surgical timing by applying ML models. They found that the main factors associated with postoperative mortality in patients who underwent early surgery (<4 weeks) included modified Marshall score on admission and preoperational modified Marshall score. Preoperational modified Marshall score, time of surgery, duration of organ failure and onset of renal failure were important predictive factors for the postoperative mortality of patients who underwent delayed surgery (≥4 weeks). Finally, the random forest model with better performance than common statistic model, was constructed to predict the surgical timing, providing good references for clinicians in developing personalized surgical plans for patients with IPN. Nevertheless, the study was limited by its retrospective,

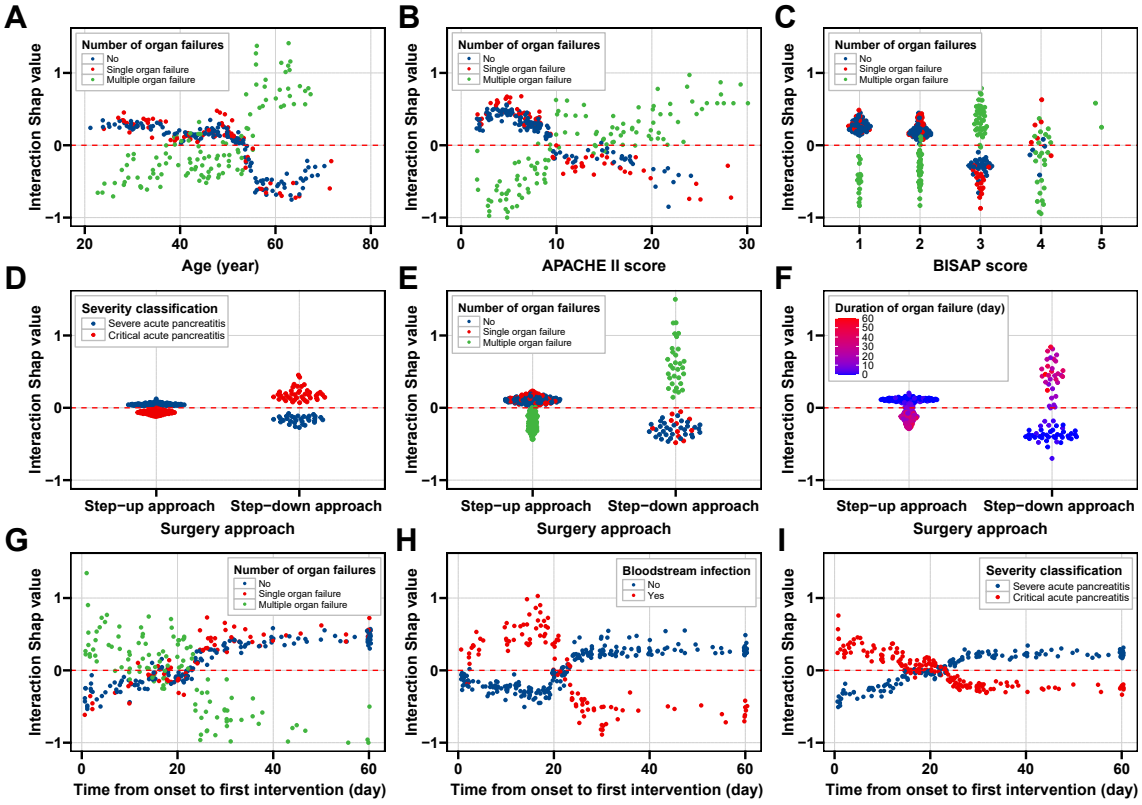

Fig. 7: SHAP dependence plot of nine feature pairs based on interaction SHAP value in the training cohort. Plots visualized nine feature pairs with relative high interaction value, including number of organ failures vs. age (A), APACHE II score (B), and BISAP score (C); surgery approach vs. severity classification (D), number of organ failures (E), and duration of organ failure (F); and time from onset to first intervention vs. number of organ failures (G), bloodstream infection (H), and severity classification (I). Plots just indicated the interaction effect of two features on death, with higher interaction SHAP value representing higher addictive death risk. SHAP, SHapley Additive exPlanations; APACHE II score, Acute Physiology and Chronic Health Examination II score; BISAP score, Bedside Index of Severity in Acute Pancreatitis score.

single-center design, lacking external validation. Moreover, its applicability was restricted to patients who underwent surgery, thus excluding those who did not undergo surgical intervention. In the present study, we applied RSF model, better compared with other 9 ML models, through a large prospective cohort to predict death and identified the key factors associated with mortality in patients with IPN. The RSF model was constructed by 10 variables (age, smoking or drinking, APACHE II score, number of organ failures, duration of organ failure, bloodstream infection, pancreatic CRE infection, time from onset to first intervention, surgery approach and hemorrhage) and performed well both in internal and external validation, with C-index of 0.863 and 0.857, respectively.

Notably, we utilized the SHAP algorithm to identify key predictors, offering a general overview of feature importance and their impact on model predictions.[25,32] The top 9 important predictors associated with increased mortality included MOF, APACHE II score ≥20, duration of organ failure ≥21 days, bloodstream infection, time from onset to first intervention <30 days, BISAP score ≥3, critically severe acute pancreatitis, age ≥ 50 years, and hemorrhage, which was consistent with the results of our previous study.[24] These results suggested persistent organ failure should be reversed within 21 days, and surgical intervention should, whenever possible, be delayed until at least 30 days from the onset. More effective treatment measures were needed to reduce mortality in patients with MOF, APACHE II score ≥20, or a BISAP score ≥3. Meanwhile, we explored interactive effects between variables, revealing the intricate relationship between two predictors and their influence on mortality.[24] For example, patients with age ≥ 50 years and MOF exhibit a positive interaction effect, indicating that patients would fare worse than expected from the additive prognostic effect of the two variables. This finding highlighted the necessity of implementing effective measures to reverse MOF in elderly IPN patients. The addictive positive interaction was further confirmed by interaction metrics in the traditional model. Similar high interactive effects

were also observed in the other eight pairs of features: (1) number of organ failures and APACHE II score; (2) number of organ failures and BISAP score; (3) number of organ failures and time from onset to first intervention; (4) number of organ failures and surgery approach; (5) severity classification and surgery approach; (6) severity classification and time from onset to first intervention; (7) duration of organ failure and surgery approach; (8) bloodstream infection and time from onset to first intervention. These interactive effects indicated that the patients would have worse prognosis when any of the above paired features were present simultaneously. Notably, the step-down approach should be avoided whenever possible in cases of critical acute pancreatitis, particularly in patients with MOF and a duration of organ failure ≥21 days.

To our knowledge, this was the first time that interactive effects between death predictors were demonstrated in IPN patients, which provided deep insights into how the RSF model made its decisions. Two free, interactive web-based Shiny apps for RSF model were constructed to provide both death prediction and explanation for individuals, enhancing its usability among clinicians. The ML model offered two significant benefits. First, it enabled clinicians to accurately predict the risk of mortality for IPN patients, allowing for early identification of high-risk patients and the implementation of targeted interventions and closer monitoring. Second, it served as a valuable tool for facilitating communication with patients and their families regarding prognosis and treatment options.

However, there were several limitations which should be acknowledged in this study. First, the model was developed using a single-center, prospective cohort over an extended time period, raising concerns about its generalizability to other centers and global populations. The prolonged study period may introduce variability in clinical management practices and data collection due to evolving guidelines and treatment strategies for IPN, potentially resulting in heterogeneous effects on clinical outcomes and limiting the applicability. Furthermore, while the two hospitals included in the study differed significantly in patients volume, both were located within the same province, limiting the model's generalizability. Second, IPN was defined based on a positive culture of pancreatic necrosis or fluid obtained during the first drainage or necrosectomy. Although the definition has a low false-positive rate, it may exclude patients with IPN successfully treated with non-surgical therapy. Third, biomarkers (e.g., gene or protein expression data) were not collected in the screening of risk factors for death prediction due to the absence of established biomarkers, although important biomarkers may enhance disease understanding and outcome prediction in various contexts. The use of antibiotics (e.g., types, duration, prophylactic or therapeutic use) was not well documented due to the complexity of antibiotic treatment for IPN. Our study focused on routinely collected, clinically accessible variables for immediate practical application. Fourth, while the combination of ML models and interpretable SHAP algorithm facilitated clinician trust and meaningful information extraction, it was crucial to acknowledge that clinical judgment cannot be entirely replaced by the model. Fifth, though ML model has shown potential in decision-making assistance, the C-index may be considered modest in comparison to models developed with advanced artificial intelligence techniques in the era of precision medicine. Future studies would focus on constructing multicenter large-scale databases and combine precision medicine techniques and advanced artificial intelligence to develop and validate high-quality models.

In conclusion, an explainable ML model for death prediction among IPN patients was feasible and effective. The final RSF model had an excellent ability for death prediction in both internal and external validation, suggesting its superior potential in guiding clinical management and improving patient outcomes. Two publicly accessible web tools generated for the optimized model facilitated its utility in clinical settings.

Contributors

CN: Conceptualization, Data curation, Project administration, Validation, Writing–Original Draft, Writing–Review & Editing.

HO: Methodology, Formal analysis, Software, Visualization, Writing–Original Draft, Writing–Review & Editing.

JX, DW, ZS, BL, DS, XH, CL, JL, LC, SZ, XL: Data curation, Methodology, Writing–Review & Editing.

FX: Conceptualization, Methodology, Supervision, Writing–Review & Editing, Project administration.

GH: Conceptualization, Methodology, Supervision, Resources, Funding acquisition, Writing–Review & Editing, Project administration.

All authors approved the final version of the manuscript. CN and GH accessed and verified the data underlying this study.

Data sharing statement

The datasets used and/or analyzed for the present study are available from the corresponding author (Gengwen Huang, huanggengwen@csu.edu.cn) on reasonable request.

Declaration of interests

All authors have no conflicts of interest to disclose.

Acknowledgements

This study was supported financially by the Natural Science Foundation of Hunan Province (2023JJ30885), Postdoctoral Fellowship Program of CPSF (GZB20230872), the Youth Science Foundation of Xiangya Hospital (2023Q13) and the Project Program of National Clinical Research Center for Geriatric Disorders of Xiangya Hospital (2021LNJJ19).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103074.

References

1 Forsmark CE, Vege SS, Wilcox CM. Acute pancreatitis. *N Engl J Med*. 2016;375(20):1972–1981. https://doi.org/10.1056/NEJMra1505202.

2 Mederos MA, Reber HA, Girgis MD. Acute pancreatitis: a review. *JAMA*. 2021;325(4):382–390. https://doi.org/10.1001/jama.2020.20317.

3 Xiao AY, Tan ML, Wu LM, et al. Global incidence and mortality of pancreatic diseases: a systematic review, meta-analysis, and meta-regression of population-based cohort studies. *Lancet Gastroenterol Hepatol*. 2016;1(1):45–55. https://doi.org/10.1016/S2468-1253(16)30004-8.

4 van Dijk SM, Hallensleben NDL, van Santvoort HC, et al. Acute pancreatitis: recent advances through randomised trials. *Gut*. 2017;66(11):2024–2032. https://doi.org/10.1136/gutjnl-2016-313595.

5 Stirling AD, Moran NR, Kelly ME, Ridgway PF, Conlon KC. The predictive value of C-reactive protein (CRP) in acute pancreatitis - is interval change in CRP an additional indicator of severity? *HPB (Oxford)*. 2017;19(10):874–880. https://doi.org/10.1016/j.hpb.2017.06.001.

6 Koutroumpakis E, Wu BU, Bakker OJ, et al. Admission hematocrit and rise in blood urea nitrogen at 24 h outperform other laboratory markers in predicting persistent organ failure and pancreatic necrosis in acute pancreatitis: a post hoc analysis of three large prospective databases. *Am J Gastroenterol*. 2015;110(12):1707–1716. https://doi.org/10.1038/ajg.2015.370.

7 Sahu B, Abbey P, Anand R, Kumar A, Tomer S, 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*. 2017;27(2):152–160. https://doi.org/10.4103/ijri.IJRI_300_16.

8 Maisonneuve P, Lowenfels AB, Lankisch PG. The harmless acute pancreatitis score (HAPS) identifies non-severe patients: a systematic review and meta-analysis. *Pancreatology*. 2021;21(8):1419–1427. https://doi.org/10.1016/j.pan.2021.09.017.

9 Podda M, Pellino G, Di Saverio S, et al. Infected pancreatic necrosis: outcomes and clinical predictors of mortality. A post hoc analysis of the MANCTRA-1 international study. *Updates Surg*. 2023;75(3):493–522. https://doi.org/10.1007/s13304-023-01488-6.

10 Moran RA, Halloran C, Guo Q, et al. Early infection is an independent risk factor for increased mortality in patients with culture-confirmed infected pancreatic necrosis. *Pancreatology*. 2022;22(1):67–73. https://doi.org/10.1016/j.pan.2021.11.003.

11 Wu D, Xiao J, Ding J, et al. Predictors of mortality and drug resistance among carbapenem-resistant enterobacteriaceae-infected pancreatic necrosis patients. *Infect Dis Ther*. 2021;10(3):1665–1676. https://doi.org/10.1007/s40121-021-00489-5.

12 Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.

13 Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nat Med*. 2023;29(7):1804–1813. https://doi.org/10.1038/s41591-023-02396-3.

14 Dupont T, Kentish-Barnes N, Pochard F, Duchesnay E, Azoulay E. Prediction of post-traumatic stress disorder in family members of ICU patients: a machine learning approach. *Intensive Care Med*. 2024;50(1):114–124. https://doi.org/10.1007/s00134-023-07288-1.

15 Gilholm P, Gibbons K, Brüningk S, et al. Machine learning to predict poor school performance in paediatric survivors of intensive care: a population-based cohort study. *Intensive Care Med*. 2023;49(7):785–795. https://doi.org/10.1007/s00134-023-07137-1.

16 Liu X, Hu P, Yeung W, et al. Illness severity assessment of older adults in critical illness using machine learning (ELDER-ICU): an international multicentre study with subgroup bias evaluation. *Lancet Digit Health*. 2023;5(10):e657–e667. https://doi.org/10.1016/S2589-7500(23)00128-0.

17 Ohbe H, Goto T, Nakamura K, Matsui H, Yasunaga H. Development and validation of early prediction models for new-onset functional impairment at hospital discharge of ICU admission. *Intensive Care Med*. 2022;48(6):679–689. https://doi.org/10.1007/s00134-022-06688-z.

18 Qiu W, Chen H, Dincer AB, Lundberg S, Kaeberlein M, Lee SI. Interpretable machine learning prediction of all-cause mortality. *Commun Med*. 2022;2:125. https://doi.org/10.1038/s43856-022-00180-x.

19 Zhou Y, Ge YT, Shi XL, et al. Machine learning predictive models for acute pancreatitis: a systematic review. *Int J Med Inform*. 2022;157:104641. https://doi.org/10.1016/j.ijmedinf.2021.104641.

20 Jin X, Ding Z, Li T, Xiong J, Tian G, Liu J. Comparison of MPL-ANN and PLS-DA models for predicting the severity of patients with acute pancreatitis: an exploratory study. *Am J Emerg Med*. 2021;44:85–91. https://doi.org/10.1016/j.ajem.2021.01.044.

21 Xu F, Chen X, Li C, et al. Prediction of multiple organ failure complicated by moderately severe or severe acute pancreatitis based on machine learning: a multicenter cohort study. *Mediators Inflamm*. 2021;2021:5525118. https://doi.org/10.1155/2021/5525118.

22 Ding N, Guo C, Li C, Zhou Y, Chai X. An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III. *BioMed Res Int*. 2021;2021:6638919. https://doi.org/10.1155/2021/6638919.

23 Lan L, Guo Q, Zhang Z, et al. Classification of infected necrotizing pancreatitis for surgery within or beyond 4 Weeks using machine learning. *Front Bioeng Biotechnol*. 2020;8:541. https://doi.org/10.3389/fbioe.2020.00541.

24 Ning C, Ouyang H, Shen D, et al. Prediction of survival in patients with infected pancreatic necrosis: a prospective cohort study. *Int J Surg*. 2024;110(2):777–787. https://doi.org/10.1097/JS9.0000000000000844.

25 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems (NIPS'17)*. Curran Associates; 2017:4765–4774.

26 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. https://doi.org/10.1136/bmj.g7594.

27 Mathew G, Agha R, Albrecht J, et al. STROCSS 2021: Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. *Int J Surg*. 2021;96:106165. https://doi.org/10.1016/j.ijsu.2021.106165.

28 Park SY, Park JE, Kim H, Park SH. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean J Radiol*. 2021;22(10):1697–1707. https://doi.org/10.3348/kjr.2021.0223.

29 Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.

30 Sternby H, Bolado F, Canaval-Zuleta HJ, et al. Determinants of severity in acute pancreatitis: a nation-wide multicenter prospective cohort study. *Ann Surg*. 2019;270(2):348–355. https://doi.org/10.1097/SLA.0000000000002766.

31 Mofidi R, Duff MD, Madhavan KK, Garden OJ, Parks RW. Identification of severe acute pancreatitis using an artificial neural network. *Surgery*. 2007;141(1):59–66. https://doi.org/10.1016/j.surg.2006.07.022.

32 Crombé A, Kataoka M. Breast cancer molecular subtype prediction: improving interpretability of complex machine-learning models based on multiparametric-MRI features using SHapley Additive exPlanations (SHAP) methodology. *Diagn Interv Imaging*. 2024;105(5):161–162. https://doi.org/10.1016/j.diii.2024.01.008.