BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Machine learning models for predicting in-hospital mortality from acute pancreatitis in intensive care unit

Shuxing Wei[1†], Hongmeng Dong[1†], Weidong Yao[2†], Ying Chen[1], Xiya Wang[1], Wenqing ji[1], Yongsheng Zhang[3*] and Shubin Guo[1*]

Abstract

Background Acute pancreatitis (AP) represents a critical medical condition where timely and precise prediction of in-hospital mortality is crucial for guiding optimal clinical management. This study focuses on the development of advanced machine learning (ML) models to accurately predict in-hospital mortality among AP patients admitted to intensive care unit (ICU).

Method Our study utilized data from three distinct sources: the Medical Information Mart for Intensive Care III (MIMIC-III), MIMIC-IV databases, and Beijing Chaoyang Hospital. We systematically developed and evaluated 11 distinct machine learning (ML) models, employing a comprehensive set of evaluation metrics to assess model performance, including the area under the curve (AUC). To enhance interpretability and identify key predictive features, we implemented Shapley Additive Explanations (SHAP) analysis for the top-performing model. Furthermore, we developed a streamlined version of the model through strategic feature reduction, followed by rigorous hyperparameter optimization (HPO) to maximize predictive performance. To facilitate clinical implementation, we designed and deployed an intuitive web-based calculator, enabling convenient access and practical application of our optimized predictive model.

Result The study analyzed 1802 AP patients, with 266 (14.8%) experiencing in-hospital mortality. A set of 27 features was utilized to construct various models, and among them, CatBoost demonstrated the highest performance in both the validation and test sets. To create a more concise model, we selected the top 13 features. After HPO, the AUC in the test set reached 0.835 (95% CI: 0.793–0.872), the AUC in the external validation from Beijing Chaoyang hospital was 0.782 (95% CI: 0.699–0.860).

Conclusion ML models have shown promising reliability in predicting in-hospital mortality among patients with AP in the ICU. Among these models, the CatBoost model exhibits superior predictive performance, providing valuable

[†] Shuxing Wei, Hongmeng Dong, and Weidong Yao contributed equally to this work.

*Correspondence:
Yongsheng Zhang
ruc_zys@126.com
Shubin Guo
shubin007@yeah.net

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

assistance to clinical practitioners in identifying high-risk patients and facilitating early interventions to enhance prognosis. The development of a compact model and a web-based calculator further enhances the convenience of using these models in clinical practice.

Keywords Acute pancreatitis, Intensive care unit, Machine learning, Mortality, Decision-making

Introduction

Acute pancreatitis (AP) represents a prevalent gastrointestinal disorder requiring hospitalization, with its incidence demonstrating a consistent upward trend over the past six decades [1]. Epidemiological data reveal an annual mortality rate ranging from 6.9 to 11.7 per million population [2], which escalates significantly to 20–40% in cases progressing to moderate or severe disease [3]. Notably, a comprehensive multicenter study conducted in Australia and New Zealand reported an 11.6% in-hospital mortality rate among AP patients admitted to intensive care units [4]. These concerning statistics underscore the critical importance of early disease recognition and timely intervention, particularly through evidence-based measures such as prompt fluid resuscitation, to improve clinical outcomes and reduce mortality in AP patients.

Machine learning (ML) represents a transformative modeling approach [5] that outperforms traditional methods like logistic regression by automatically uncovering complex variable relationships in large datasets. This advanced capability enables more accurate predictions through optimal feature selection and pattern recognition. Particularly valuable in healthcare research [6]. ML excels in processing complex, high-dimensional data, making it indispensable for predictive tasks including disease diagnosis, complication prediction, and medication anomaly detection.

This study aims to develop a predictive model for mortality risk in AP patients, facilitating early identification of high-risk cases. By identifying modifiable risk factors associated with disease progression and mortality, our research seeks to inform the development of targeted therapeutic strategies to prevent disease progression from mild to severe stages. This dual approach not only promises to improve patient outcomes but also has the potential to significantly reduce healthcare costs [7].

Methods

Data source

We extracted admission data for AP patients from the Medical Information Mart for Intensive Care III (MIMIC-III) [8] and MIMIC-IV [9] databases, using International Classification of Diseases codes (ICD-9: 577.0; ICD-10: K85-K85.92) for case identification. The MIMIC database represents a comprehensive clinical repository encompassing detailed patient care information, including demographic data, physiological monitoring parameters, laboratory results, microbiological findings, fluid balance records, pharmacological interventions, hospitalization duration, and clinical outcomes. The study protocol received Institutional Review Board approval (Certification No: 47937607), and all database access procedures complied with ethical standards for human subject protection. The external validation dataset was obtained from the Emergency Intensive Care Unit (EICU) at Beijing Chaoyang Hospital, which is affiliated with Capital Medical University, China.

Patients enrollment and data collection

The study population comprised adult patients with ICU admissions for acute pancreatitis. We established the following exclusion criteria: (1) ICU stays shorter than 24 h, (2) admissions for recurrent AP episodes, and (3) cases with more than 30% missing data for the dependent variable. For model development, we implemented a stratified random sampling approach, allocating 80% of the MIMIC-IV dataset to the training cohort and reserving 20% for validation, ensuring proper representation of clinical characteristics across both cohorts. An set from MIMIC-III was used for test, and external validation was conducted using data between January 2020 and June 2024 from EICU of Beijing Chaoyang Hospital to assess the model's generalizability.

Data preprocessing included outlier detection using the Z-score method, where data points with a Z-score exceeding 3 were considered outliers and removed to minimize their impact on model performance. Missing values were handled using multiple imputation with the *IterativeImputer* class. This method iteratively estimates missing values through regression models, refining imputations until convergence or reaching the maximum iteration limit ($max\_iter = 10$), thereby reducing bias introduced by missing data. To ensure consistency in scale and improve model performance, continuous variables were standardized using the *StandardScaler* method from the *sklearn.preprocessing* library, transforming data to have a mean of 0 and a standard deviation of 1.

The study incorporated the following information: (1) Demographic characteristics, such as sex and age. (2) Comorbidities, including type 2 diabetes mellitus (T2DM), hyperlipidemia, hypertension, atrial fibrillation (AF), acute myocardial infarction (AMI), heart failure (HF), renal failure (RF), and sepsis. (3) Vital signs, encompassing heart rate (HR), temperature, oxygen saturation (SpO2), systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP). (4) Laboratory parameters, such as platelets, blood urea nitrogen (BUN), calcium, potassium, sodium, creatinine, glucose, hematocrit, hemoglobin, white blood cell (WBC) counts, alanine aminotransferase (ALT), aspartate aminotransferase (AST), international normalized ratio (INR), and lactate. (5) Treatment measures, including continuous renal replacement therapy (CRRT) and mechanical ventilation (MV). Regarding renal disease, it encompasses acute kidney injury, chronic renal insufficiency, and end-stage renal disease. MV includes both invasive and non-invasive ventilation methods.A total of 32 variables were considered in the analysis. In the case of variables that were measured multiple times, only the initial measurement was included in the analysis.

Model construction and evaluation

Our analytical approach commenced with LASSO regression for optimal feature selection, followed by the development of eleven distinct machine learning models: logistic regression, K-nearest neighbors (KNN), decision tree, random forest (RF), support vector machine (SVM), XGBoost, AdaBoost, gradient boosting decision trees (GBDT), multilayer perceptron (MLP), LightGBM, and CatBoost. To enhance model reliability and generalizability, we implemented a rigorous k-fold cross-validation framework during both model training and parameter optimization. This validation strategy systematically partitions the dataset into k subsets, enabling comprehensive performance evaluation while effectively mitigating overfitting risks and reducing dataset-specific bias.

Among these 11 models, the traditional logistic regression model [10] has the advantage of being able to visualize the results using techniques such as column charts. KNN regression is a simple yet practical method based on feature similarity [11]. Predicting the labels of query samples involves two steps. Firstly, the distances between the query sample and the samples in the train set are calculated to determine its k nearest neighbors. Then, the average or median of these neighboring labels is taken as the final prediction. Decision tree [12] is a highly versatile machine learning model that can be used for both regression and classification tasks. A decision tree is a tree-like structure where each internal node represents a judgment on a specific attribute, each branch represents an output based on that judgment, and each leaf node represents a classification outcome. Random forest [13] is an ensemble of multiple individual decision tree models. SVM [14] is a fast and reliable classification algorithm that performs exceptionally well when the data volume is limited. XGBoost [15] is an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. It implements machine learning algorithms within the Gradient Boosting framework. AdaBoost [16] is a relatively recent algorithm chosen for its strong theoretical foundation, simplicity of implementation, transparency in feature selection, and performance in discriminative tasks. This algorithm simultaneously selects and combines relevant features from the feature set during the training process of each individual classifier, thus avoiding the separate feature selection process commonly seen in other classification methods. GBDT [17] is an iterative decision tree algorithm composed of multiple decision trees, where the conclusions of all trees are accumulated to produce the final answer. It can optimize for different loss functions and provides various options for hyperparameter tuning, making it highly flexible in function fitting.

MLP is a neural network model. The most typical MLP consists of three layers: the input layer, the hidden layer(s), and the output layer [18]. Each layer has several neurons that are interconnected through nodes. The neurons in the hidden layer apply an activation function to the weighted inputs and propagate the results to the nodes in the next layer. The training process involves iteratively updating the connection weights through backpropagation. LightGBM [19] is another variant of the gradient boosting tree algorithm that has gained popularity recently. It incorporates multiple new techniques to jointly optimize computation speed, memory usage, and prediction performance. CatBoost [20] is the third algorithm based on GBDT, following XGBoost and LightGBM. Since its debut in late 2018, researchers have successfully utilized CatBoost in machine learning research involving large-scale data. It is particularly well-suited for machine learning tasks involving classification and heterogeneous data.

Following model development, we conducted comprehensive performance evaluation using independent test and validation datasets. Model performance was assessed through multiple metrics, including the receiver operating characteristic (ROC) curve, accuracy, positive predictive value (PPV), negative predictive value (NPV), F1-score, sensitivity, and specificity. To optimize clinical utility, we specifically applied feature reduction and hyperparameter optimization (HPO) techniques to the best-performing model, subsequently comparing the streamlined version's performance against its original full-featured counterpart. Calibration curves were employed to evaluate model fit and prediction reliability. For clinical benchmarking, we implemented a logistic regression model based on the acute physiology and chronic health evaluation (APACHE) II scoring system, enabling direct comparison of area under the curve (AUC) values with our model to demonstrate clinical relevance and potential superiority.

Statistical analysis

Categorical variables were presented as frequencies and percentages, with between-group comparisons performed using chi-square tests. Statistical significance for quantitative data was evaluated using Student's t-test or the nonparametric Wilcoxon test, depending on the results of the normality test; the chi-square test was used for qualitative data.

Following model evaluation, we employed the SHapley Additive exPlanations (SHAP) method to quantify and interpret feature importance in our optimal model. Rooted in cooperative game theory, SHAP provides a unified framework for machine learning interpretability by constructing an additive feature attribution model. This approach assigns each feature an importance value while maintaining both local precision (individual prediction explanation) and global consistency (overall model interpretation). The SHAP values offer dual advantages: quantitatively ranking feature contributions and qualitatively revealing the directionality (positive or negative impact) of each feature's influence on model predictions [21]. We implemented feature reduction on the optimal model, followed by HPO to enhance the streamlined model's performance. HPO employs algorithmic approaches to identify optimal hyperparameter configurations, eliminating reliance on manual tuning. This automated process systematically generates multiple hyperparameter sets, iteratively trains models, and refines parameter selection based on performance metrics. To facilitate clinical implementation, we developed an intuitive web-based calculator. All statistical analyses were conducted using R (version 4.1.2) and Python (version 3.10), with statistical significance defined as $P < 0.05$.

Result

Characteristics of the cohort

Our study included 1,802 AP patients across all cohorts, with detailed enrollment illustrated in Fig. 1. The MIMIC-IV cohort comprised 812 patients, demonstrating a 12.6% in-hospital mortality rate ($n = 102$). Similarly, the test cohort included 799 patients with a 15.8% mortality rate ($n = 126$). The external validation cohort consisted of 199 patients, showing a higher mortality rate of 19.1% ($n = 38$). Comprehensive patient characteristics are documented in Supplementary Tables 1–3.

Model construction and evaluation

We included 32 variables in the LASSO regression for feature selection. A total of 27 variables—ALT, AST, BUN, calcium, creatinine, DBP, hematocrit, hemoglobin, HR, INR, lactate, sodium, platelet count, SBP, SpO_2, temperature, WBC, gender, age, hypertension, AMI, sepsis, T2DM, hyperlipidemia, CRRT, RF, and MV—were ultimately selected for model construction. The variable importance is presented in Supplementary Fig. 1.

Following the development of 11 machine learning models, we initially assessed their performance on the validation set. As summarized in Table 1, the CatBoost model demonstrated superior predictive capabilities, achieving the highest accuracy (90.80%), PPV (0.6250), sensitivity (0.8022), and specificity (0.9080). While its NPV(0.9226) was marginally lower than the Decision Tree model (0.9357), CatBoost exhibited optimal

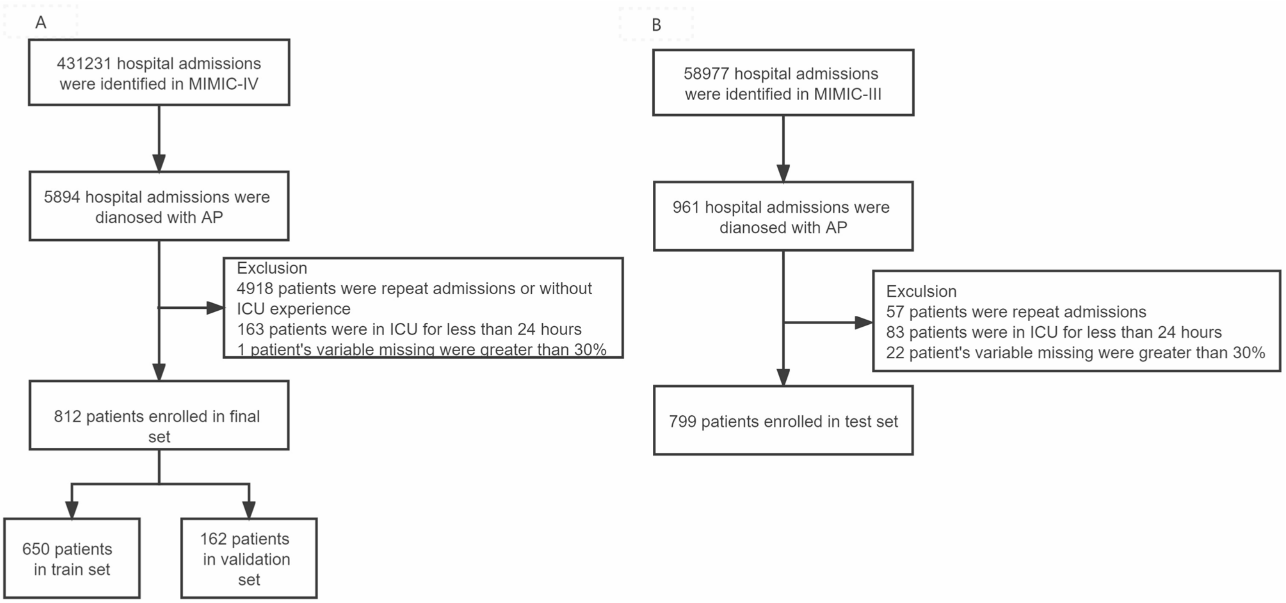
A
431231 hospital admissions were identified in MIMIC-IV
5894 hospital admissions were dianosed with AP
Exclusion
4918 patients were repeat admissions or without ICU experience
163 patients were in ICU for less than 24 hours
1 patient's variable missing were greater than 30%
812 patients enrolled in final set
650 patients in train set
162 patients in validation set
B
58977 hospital admissions were identified in MIMIC-III
961 hospital admissions were dianosed with AP
Exculsion
57 patients were repeat admissions
83 patients were in ICU for less than 24 hours
22 patient's variable missing were greater than 30%
799 patients enrolled in test set

Fig. 1 Flowchart of the selection process of patients. (**A**) Flowchart of MIMIC-IV; (**B**) Flowchart of MIMIC-III. *MIMIC, Medical Information Mart for Intensive Care; AP, acute pancreatitis; ICU, intensive care unit*

Table 1 Performance of 11 models in validation set

Model	Accuracy (%)	PPV	NPV	F1_score	Sensitivity	Specificity
Logistic_Regression	88.96 (86.78–97.56)	0.4000 (0.3805–0.4195)	0.9051 (0.8886–0.9216)	0.8616 (0.8206–0.9026)	0.6690 (0.5432–0.7754)	0.8910 (0.7742–0.9532)
KNN	89.57 (86.25–94.27)	0.5000 (0.4922–0.5078)	0.9006 (0.8578–0.9434)	0.8571 (0.8436–0.8706)	0.7657 (0.7442–0.7872)	0.8950 (0.8691–0.9209)
Decision_Tree	85.28 (81.09–93.58)	0.3600 (0.3452–0.3748)	0.9357 (0.8915–0.9799)	0.8647 (0.8541–0.8753)	0.7528 (0.7270–0.7786)	0.8530 (0.8440–0.8620)
Random_Forest	90.18 (88.33–94.78)	0.5714 (0.5604–0.5824)	0.9167 (0.8877–0.9457)	0.8830 (0.8695–0.8965)	0.7238 (0.6878–0.7598)	0.9020 (0.8649–0.9391)
SVM	87.73 (84.85–96.77)	0.3636 (0.3554–0.3718)	0.9145 (0.8782–0.9508)	0.8654 (0.8361–0.8947)	0.5799 (0.5590–0.6008)	0.8770 (0.8560–0.8980)
XGBoost	90.18 (89.13–92.47)	0.5556 (0.5289–0.5823)	0.9221 (0.9114–0.9328)	0.8880 (0.8583–0.9177)	0.6683 (0.6409–0.6957)	0.9020 (0.8623–0.9417)
AdaBoost	90.41 (88.30–93.10)	0.5882 (0.5730–0.6034)	0.9121 (0.8713–0.9529)	0.8741 (0.8448–0.9034)	0.6441 (0.5592–0.6990)	0.9040 (0.8791–0.9289)
GBDT	88.96 (85.32–97.92)	0.4545 (0.4427–0.4663)	0.9211 (0.8997–0.9425)	0.8789 (0.8541–0.9037)	0.7796 (0.7193–0.8499)	0.8900 (0.8483–0.9317)
MLP	85.89 (84.86–94.88)	0.3125 (0.2983–0.3267)	0.9184 (0.8857–0.9511)	0.8570 (0.8169–0.8971)	0.5889 (0.5149–0.6529)	0.8590 (0.8211–0.8969)
LightGBM	89.57 (86.16–98.96)	0.5000 (0.4902–0.5098)	0.9216 (0.8784–0.9648)	0.8834 (0.8687–0.8981)	0.6857 (0.6213–0.7601)	0.8960 (0.8617–0.9303)
CatBoost	90.8000 (88.21–95.31)	0.6250 (0.6018–0.6482)	0.9226 (0.9075–0.9377)	0.8928 (0.8700-0.9156)	0.8022 (0.7624–0.8420)	0.9080 (0.8799–0.9361)

KNN, K-nearest neighbors; SVM, support vector machine, GBDT, gradient boosting decision trees, MLP, multilayer perceptron; PPV, positive predictive value, NPV, negative predictive value

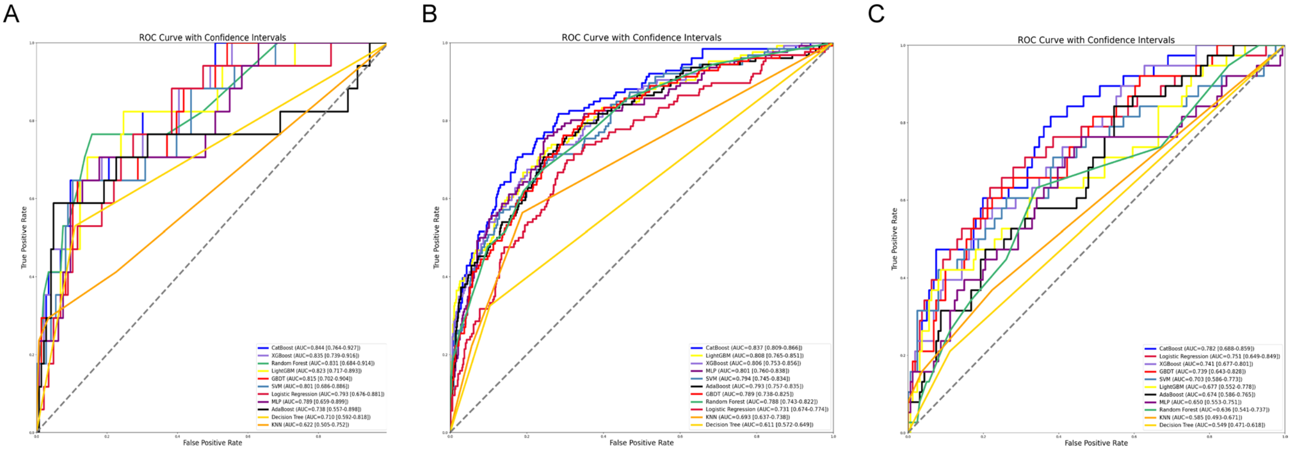

Fig. 2 ROC curves of 11 machine learning models in the validation set, test set, and external validation set, illustrating their classification performance. (**A**) ROC curves in the validation set; (**B**) ROC curves in the test set; (**C**) ROC curves in the external validation set. Higher AUC values indicate better discriminative ability of the models. *ROC, receiver operating characteristic; AUC, area under the curve; KNN, K-nearest neighbors; SVM, support vector machine; GBDT, gradient boosting decision trees; MLP, multilayer perceptron*

model fitting with the highest AUC of 0.844 (95% confidence internal [CI]:0.764–0.927) (Fig. 2A). Therefore, we selected the CatBoost model as the optimal model for further study. Notably, its predictive performance significantly surpassed the conventional APACHE II scoring system (AUC = 0.64; 95%CI: 0.48–0.80) (Fig. 3A). These findings were consistently replicated in the test set, where CatBoost maintained its superior performance with an AUC of 0.837 (95%CI: 0.809–0.866) (Fig. 2B), outperforming both alternative models and the APACHE II system (AUC = 0.55; 95%CI: 0.48–0.63) (Fig. 3B). External validation further confirmed CatBoost's robust predictive capability, achieving an AUC of 0.782 (95%CI: 0.688–0.859) (Fig. 2C), significantly higher than the APACHE II system's performance (AUC = 0.53; 95%CI:0.42–0.64) (Fig. 3C).

Model visualization

Figure 4 presents the impact distribution of the top 20 features in the CatBoost model, as evaluated using SHAP values. Among these, MV was the most influential, followed by BUN, age, CRRT, sepsis, lactate, SpO_2, INR,

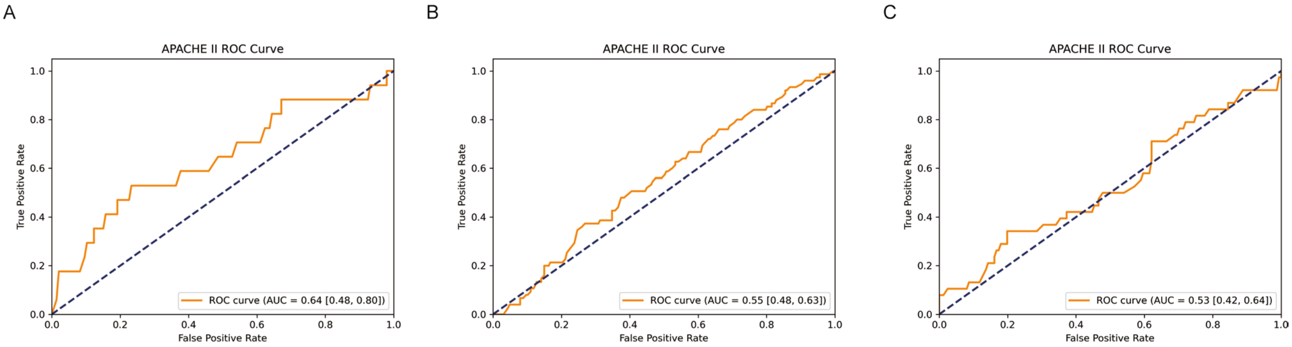

Fig. 3 ROC curves of APACHE II score in the validation set, test set, and external validation set, illustrating their classification performance. (**A**) ROC curves in the validation set; (**B**) ROC curves in the test set; (**C**) ROC curves in external validation set. *ROC*, *receiver operating characteristic*

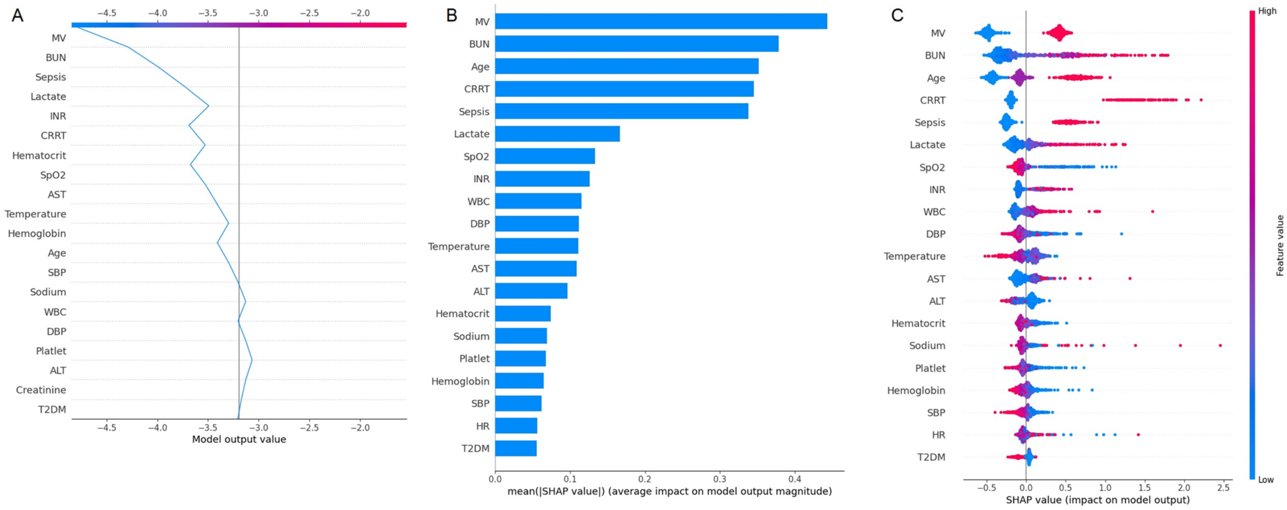

Fig. 4 SHAP analysis demonstrating the contribution of the top 20 features to the CatBoost models. (**A**) The SHAP values for a random sample, showing how individual features influence model predictions; (**B**) Feature importance ranking in the CatBoost model, highlighting the most influential variables; (**C**) SHAP summary plot displaying the impact of each feature across all patients in the CatBoost model, where larger absolute SHAP values indicate stronger influence on predictions. *SHAP*, *Shapley additive explanations; MV*, *mechanical ventilation; BUN*, *blood urea nitrogen; CRRT*, *continuous renal replacement therapy; SpO2*, *oxygen saturation; INR*, *international normalized ratio; WBC*, *white blood cell count; DBP*, *diastolic blood pressure; AST*, *aspartate aminotransferase; ALT*, *alanine aminotransferase; SBP*, *systolic blood pressure; HF*, *heart failure; T2DM*, *type 2 diabetes mellitus*

WBC, DBP, temperature, AST, ALT, hematocrit, sodium, platelet count, hemoglobin, SBP, HR, and T2DM. These features were critical in shaping the model's predictions. To ensure reliable SHAP value estimation, we employed shap.TreeExplainer, which is specifically designed for tree-based models like CatBoost. SHAP values were computed using the entire training dataset. However, to balance computational efficiency with interpretability, we selected a representative subset of 100 samples from the training set as the background dataset for shap.decision_plot. This approach aligns with standard SHAP methodology, where a well-chosen background dataset maintains computational feasibility while preserving model interpretability. The expected value of the model's predictions, derived from shap.TreeExplainer, served as the baseline reference for SHAP force plots. The summary plots in Fig. 4 display the mean absolute SHAP values for each feature, while individual SHAP dependence plots (Supplementary Fig. 2) offer deeper insights into how variations in key features (e.g., MV, BUN, age, CRRT, sepsis, lactate, SpO_2, INR, WBC, and DBP) influence model predictions. Collectively, these visualizations enhance model interpretability and provide an intuitive understanding of feature contributions to in-hospital mortality risk in AP patients.

Model simplification and optimization

To enhance clinical applicability, we developed a streamlined version of the model by selecting the 13 most influential variables based on SHAP importance rankings, using a threshold of mean absolute SHAP value >0.1. This feature-reduced model was further optimized with the Optuna framework. Through 50 iterative trials, we identified the optimal hyperparameter configuration, with the optimization process visualized in Fig. 5A. The final hyperparameter settings are summarized in Table 2,

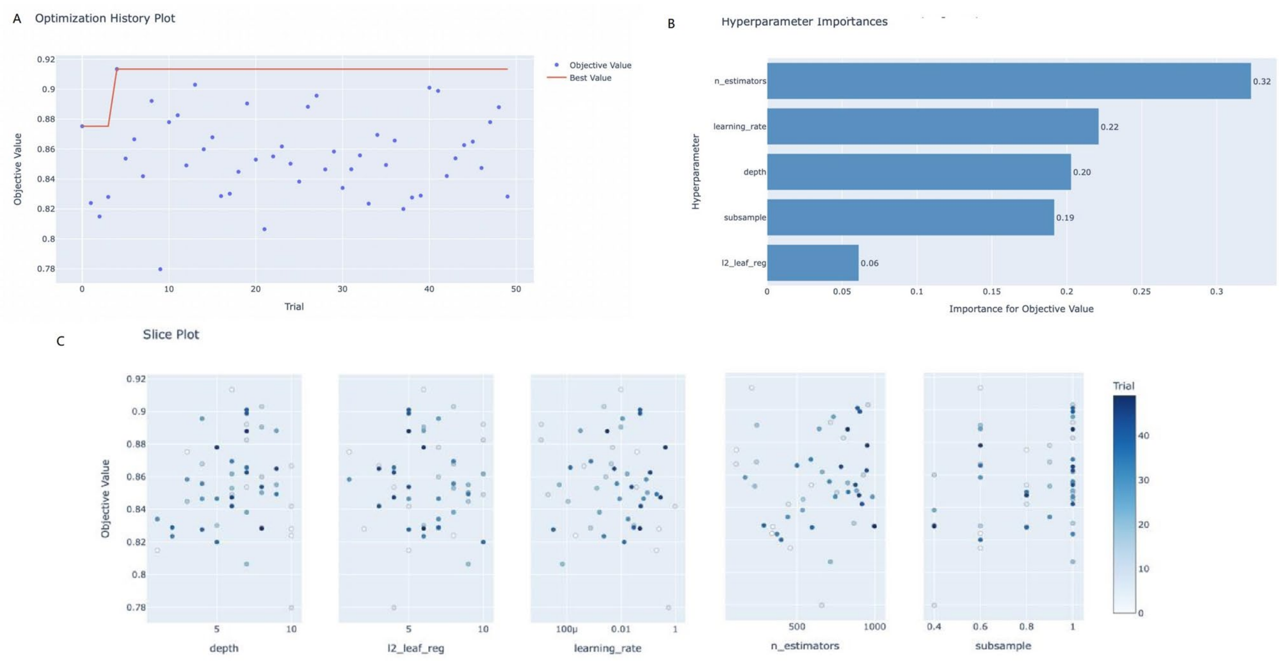

Fig. 5 Hyperparameters optimization. (**A**) Optimization history with different trials; (**B**) The importance of different hyperparameters; (**C**) The performance of single hyperparameter

Table 2 The final settings of the hyperparameters

Hyperparameters	Best Values
n_estimators	208
learning_rate	0.00985
depth	6
subsample	0.6
l2_leaf_reg	6

along with the relative importance ranking of different hyperparameters for model performance (Fig. 5B). Additionally, Fig. 5C illustrates the individual impact of key hyperparameters on model performance, offering valuable insights for parameter tuning.

Optimized model model evaluation

Comprehensive model evaluation demonstrated consistent predictive performance across multiple datasets. In the validation cohort, the full model achieved an AUC of 0.844 (95%CI: 0.748–0.918), with the compact model showing comparable performance (AUC = 0.836; 95% CI: 0.743–0.905). Following HPO, the compact model exhibited enhanced predictive capability, reaching an AUC of 0.851 (95% CI: 0.781–0.922) (Fig. 6A).

Analysis of the test set (Table 3) revealed strong performance across all model configurations. The full CatBoost model achieved 86.11% accuracy (PPV = 0.8947, NPV = 0.8603, F1_score = 0.8149, sensitivity = 0.7511,

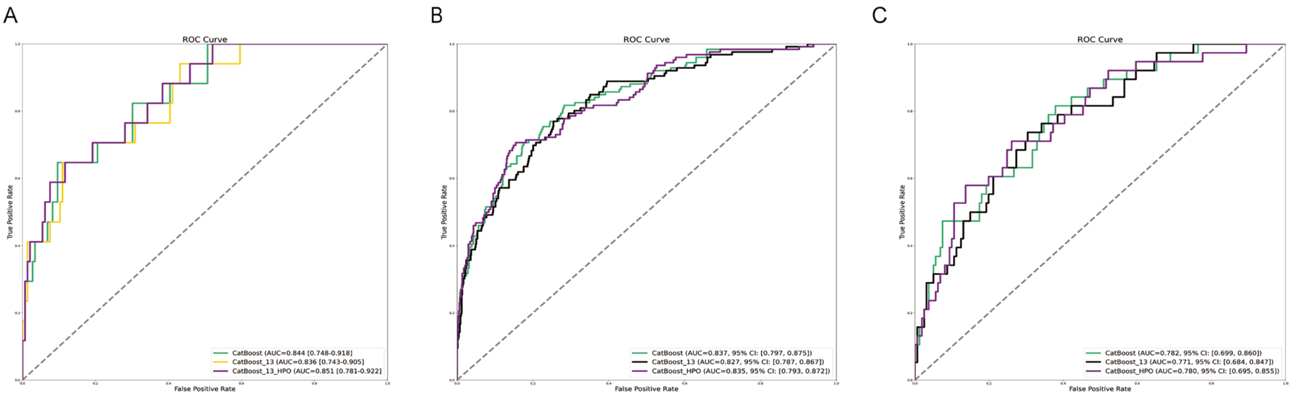

Fig. 6 ROC curves comparing the performance of CatBoost, CatBoost_13, and CatBoost_13_HPO models in the validation and test sets. (**A**) ROC curve in the validation set; (**B**) ROC curve in the test set; (**C**) ROC curve in the external validation set. These curves illustrate the impact of feature selection (CatBoost_13) and hyperparameter optimization (CatBoost_13_HPO) on model performance. *ROC, receiver operating characteristic; AUC, area under the curve*

Table 3 Performance of catboost, CatBoost_13 and CatBoost_13_HPO models in test set

Model	Accuracy (%)	PPV	NPV	F1_score	Sensitivity	Specificity
CatBoost	86.11 (84.53–89.84)	0.8947 (0.8832–0.9062)	0.8603 (0.8200-0.9006)	0.8149 (0.7742–0.8556)	0.7511 (0.6812–0.8010)	0.8603 (0.8401–0.8805)
CatBoost_13	84.98 (81.16-88.00)	0.7500 (0.7331–0.7669)	0.8513 (0.8348–0.8678)	0.7936 (0.7836–0.8036)	0.7398 (0.6723–0.7857)	0.8513 (0.8353–0.8673)
CatBoost_13_HPO	85.98 (82.34–89.99)	0.7917 (0.7728–0.8106)	0.8619 (0.8490–0.8748)	0.8171 (0.7970–0.8372)	0.7598 (0.7968–0.8228)	0.8619 (0.8263–0.8975)

PPV, positive predictive value, NPV, negative predictive value

Table 4 Performance of catboost, CatBoost_13 and CatBoost_13_HPO models in external validation set

Model	Accuracy (%)	PPV	NPV	F1_score	Sensitivity	Specificity
CatBoost	82.41 (79.40-86.77)	1.0000 (0.9880-1.0000)	0.8214 (0.7969–0.8459)	0.7577 (0.7484–0.7670)	0.6623 (0.5823–0.7452)	0.8214 (0.8093–0.8335)
CatBoost_13	82.41 (78.81–85.31)	0.8000 (0.7825–0.8175)	0.8247 (0.7934–0.8560)	0.7648 (0.7565–0.7731)	0.6241 (0.5878–0.7604)	0.8247 (0.8134–0.8360)
CatBoost_13_HPO	80.90 (71.63–90.12)	1.0000 (0.9839-1.0000)	0.8090 (0.7865–0.8315)	0.7236 (0.7032–0.7440)	0.6331 (0.5554–0.6711)	0.8090 (0.7972–0.8208)

PPV, positive predictive value, NPV, negative predictive value.

specificity = 0.8603), while the simplified CatBoost_13 model maintained 84.98% accuracy (PPV = 0.7500, NPV = 0.8513, F1_score = 0.7936, sensitivity = 0.7398, specificity = 0.8513). The optimized CatBoost_13_HPO model demonstrated further improvement with 85.98% accuracy (PPV = 0.7917, NPV = 0.8619, F1_score = 0.8171, sensitivity = 0.7598, specificity = 0.8619). Corresponding AUC values were 0.837 (95%CI: 0.797–0.875), 0.827 (95%CI: 0.787–0.867), and 0.835 (95%CI: 0.793–0.872), respectively (Fig. 6B).

External validation confirmed these findings (Fig. 6C), with the full model (AUC = 0.782; 95%CI: 0.699–0.860), compact model (AUC = 0.771; 95%CI: 0.684–0.847), and HPO-optimized compact model (AUC = 0.780; 95%CI: 0.695–0.855) all demonstrating robust performance. Detailed metrics (Table 4) showed consistent across models, with minimal performance degradation in the simplified and optimized versions.

Calibration analysis (Figs. 7 and 8) revealed superior agreement between predicted and observed outcomes for the CatBoost_13_HPO model compared to traditional approaches (logistic regression, XGBoost, and random forest), establishing its reliability across both development and external validation cohorts.

Model application

Finally, we have developed a web-based interactive program using Gradio, a Python framework that simplifies the demonstration of machine learning models, for everyone to utilize. This program is built upon 13 features to predict hospital outcomes of AP patients in ICU and provide corresponding probabilities (Supplementary Fig. 3). The main code for this program is available on the Hugging Face platform (https://huggingface.co/spaces/zysnathan/acute_pancreatitis/blob/main/model.py).

To effectively use the model in an ICU setting, healthcare professionals can input the relevant clinical variables of a patient as soon as they are available after ICU admission (such as MV, BUN, age, CRRT, sepsis, etc.). The model will then output a mortality risk score, which can be used in conjunction with clinical judgment to prioritize patients for early interventions such as fluid resuscitation or specific therapies aimed at improving outcomes. In practice, this model can assist in the identification of patients who may benefit from closer monitoring or more aggressive treatment, thereby potentially improving patient management and outcomes.

Discussion

Predictive models outperform clinical judgment alone in risk stratification and decision-making, machine learning models are being used to predict the prognosis of various diseases [22–25]. While traditional scoring systems such as APACHE II score and SOFA score are valuable for assessing the severity of illness [26, 27], machine learning models have the advantage of incorporating a broader range of dynamic and complex data, leading to more accurate and personalized predictions [28]. In this study, our main objective is to build a machine learning model to predict in-hospital outcomes of patients with AP who have had an ICU experience. This is crucial as although there have been several predictive models for AP, there is a scarcity of machine learning models specifically predicting the mortality rate of AP. Furthermore, many of the previous models have solely relied on neural network methods for prediction. For example, some studies have utilized deep learning models to predict the mortality of AP patients, which often showed promising results in terms of model accuracy, but tended to have limitations in interpretability and feature selection [29, 30]. Unlike

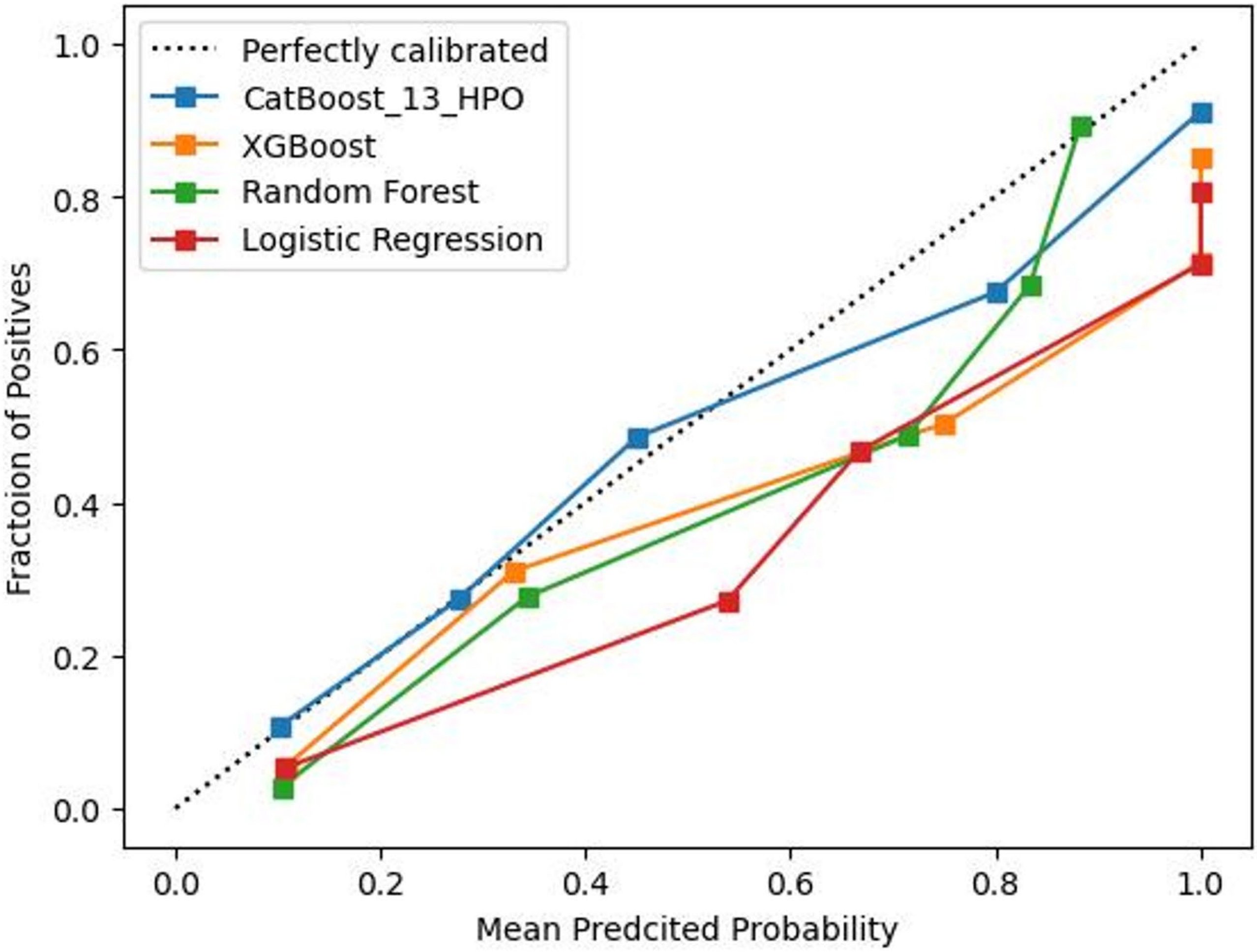

Fig. 7 The calibration curves of CatBoost_13_HPO and other models

these previous studies, our approach involves a variety of machine learning techniques, including RF and CatBoost, which provides more interpretable results, allowing us to identify key predictive features. While we primarily relied on the MIMIC-III and MIMIC-IV databases for our analysis, it is worth noting that comparing these findings to single cohort studies could offer valuable insights. Single cohort studies often allow for a more focused investigation of a specific population, which might yield results that are highly applicable to certain healthcare settings. However, the large-scale, diverse patient data available in the MIMIC databases allows for more robust model training and generalization, improving the applicability of our findings across different settings.

Early identification of the risk of death in patients admitted due to AP can improve patient outcomes significantly. By detecting high-risk patients early, healthcare providers can initiate timely and targeted interventions such as aggressive fluid resuscitation, closer monitoring, and other tailored treatments, potentially reducing mortality. Identifying risk factors for in-hospital mortality in AP patients helps prioritize high-risk individuals, thus improving clinical decision-making and resource allocation. With the advancement of machine learning algorithms, the number of predictive variables that can be handled has greatly expanded, enabling the development of optimized models compared to traditional methods. With such a model, pancreatic specialists can be alerted when a patient is admitted to the ICU with a high risk of mortality, facilitating early intervention that can improve patient outcomes.

Based on 27 clinical variables measured after ICU admission, we developed and validated 11 machine learning models to predict in-hospital outcomes for AP patients. Feature importance analysis of the best-performing model identified 13 key predictors: MV, BUN, age, CRRT, sepsis, lactate, SpO_2, INR, WBC, DBP, temperature, AST, and ALT. Using these features, we constructed a streamlined model and further optimized it through hyperparameter tuning, which maintained strong predictive performance. MV emerged as the most influential variable, reflecting its critical role in

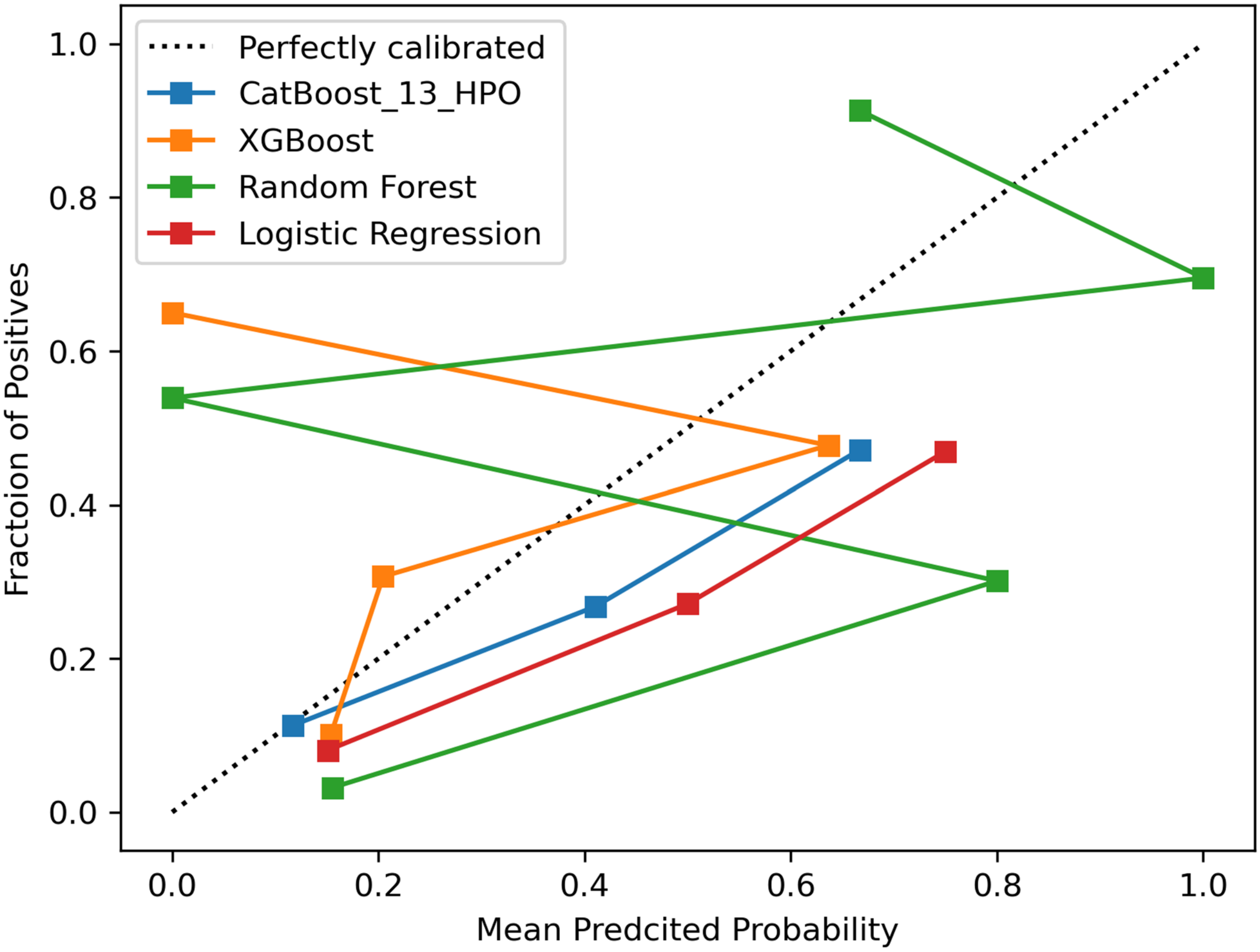

Fig. 8 The calibration curves of external validation set

assessing disease severity, as all patients in this study had ICU experience. The need for MV is often indicative of severe illness. BUN ranked second in importance, serving as an early predictor of AP outcomes, with its early measurement providing valuable prognostic insight [31, 32]. Age was the third most important predictor, consistent with prior studies showing significantly higher mortality rates among elderly AP patients [33]. CRRT was also identified as a key risk factor, aligning with previous research indicating its association with increased mortality in ICU patients with AP [34]. Among complications, sepsis had the most substantial impact on prognosis. It often originates from sterile inflammation and can escalate to severe local or systemic infections or septicemia. Sepsis remains the leading cause of death in critically ill AP patients in the ICU [35]. While these modifiable factors are crucial for risk assessment, the model's predictions should be interpreted alongside clinical judgment to guide treatment decisions effectively.

A key strength of our study is the seamless integration of model interpretation with automated feature selection to enhance both transparency and predictive performance. Unlike conventional approaches where model interpretation is often a post hoc analysis, we incorporated SHAP values and feature importance rankings throughout the feature selection and model optimization process. By using LASSO to identify the top 27 predictive variables and further refining the model to a concise 13-variable subset, we ensured both interpretability and efficiency. The interplay between model development and interpretation was further reinforced by directly incorporating SHAP analysis into feature selection, as demonstrated in prior studies such as Ikemura et al. [36]. This approach not only improves model transparency but also simplifies its structure, enhancing clinical applicability. Additionally, we developed a web-based risk calculator that visualizes mortality risk for AP patients, enabling clinicians to intuitively understand key predictive factors and the decision-making process. This practical implementation bridges automated model interpretation with real-world clinical practice, supporting personalized treatment strategies.

Despite these strengths, our study has several limitations. First, although we utilized data from MIMIC-III and MIMIC-IV, these datasets are retrospective in nature, which may introduce selection biases due to factors such as demographic variations and patient selection criteria. For instance, the MIMIC databases primarily consist of ICU patients from a specific set of hospitals, which may not fully represent the broader population, potentially limiting the generalizability of our findings to other healthcare settings. Second, despite the high quality of the MIMIC database, there are still missing data and input errors (e.g., missing values for amylase), which were handled using appropriate imputation techniques. However, we acknowledge that excluding patients with missing dependent variables exceeding 30% may introduce potential survivorship bias, as this exclusion could disproportionately affect certain patient subgroups, leading to a less representative study population. This exclusion may limit the external validity of our findings, as those with missing data may differ systematically from those with complete data. Lastly, this study focuses only on in-hospital mortality in AP patients, while other important prognostic indicators, such as long-term post-discharge mortality, require further investigation. To enhance the robustness and generalizability of our findings, future studies should consider incorporating prospective multi-center datasets that account for demographic variations and address potential confounders through methods such as propensity score matching. Additionally, further advancements in automated machine learning with integrated interpretation should be explored to refine predictive models and improve clinical applicability.

Conclusion

ML models are reliable tools for predicting in-hospital mortality rates of AP patients. In this study, the CatBoost model demonstrated the best predictive performance, helping clinical doctors identify high-risk patients to improve prognosis. Furthermore, optimized compact models and network-based calculators further enhance clinical usability.

Abbreviations

AP	Acute pancreatitis
ML	Machine learning
MIMIC	Medical Information Mart for Intensive Care
ICD	International Classification of Diseases codes
EICU	Emergency intensive care unit
T2DM	Type 2 diabetes mellitus
AF	Atrial fibrillation
AMI	Acute myocardial infarction
HF	Heart failure
RF	Renal failure
HR	Heart rate
SpO2	Oxygen saturation
SBP	Systolic blood pressure
DBP	Diastolic blood pressure
MAP	Mean arterial pressure
BUN	Blood urea nitrogen
WBC	White blood cell
ALT	Alanine aminotransferase
AST	Cytochrome P450 3A5
EPHA5	Aspartate aminotransferase
INR	International normalized ratio
CRRT	Continuous renal replacement therapy
MV	Mechanical ventilation
KNN	K-nearest neighbors
RF	Random forest
SVM	Support vector machine
GBDT	Gradient boosting decision trees
MLP	Multilayer perceptron
ROC	Receiver operating characteristic
PPV	Positive predictive value
NPV	Negative predictive value
HPO	Hyperparameter optimization
APACHE	Acute physiology and chronic health evaluation
AUC	Area under the curve
SHAP	SHapley Additive exPlanations
CI	Confidence internal

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03033-4.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

SW and HD designed the study. SW and DY wrote the manuscript. SW, HD, YC, WJ, and XW collected, analyzed and interpreted the data. YZ and SG critically reviewed, edited and approved the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Beijing Municipal Science and Technology Commission under Grant number [Z211100002921061; Z191100006619121].

Data availability

As the use of the MIMIC database requires permission, please contact us if necessary.Data will be available upon reasonable request (wsx2024@yeah.net).

Declarations

Ethics approval and consent to participate

The institutional review boards (IRB) of Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC) approved using the MIMIC-IV database (Certification No: 47937607). None of the projects required collecting individual informed consent because the protected information was de-identified.

Consent for publication

Not Applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Emergency Medicine Clinical Research Center, Beijing Key Laboratory of Cardiopulmonary Cerebral Resuscitation, Beijing Chaoyang Hospital, Affiliated to Capital Medical University, Beijing 100020, China
[2]Department of Anesthesiology, Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China

[3]Department of Health Management, Shandong Engineering Laboratory of Health Management, Institute of Health Management, the First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Jinan, China

Received: 11 April 2024 / Accepted: 14 May 2025
Published online: 27 May 2025

References

1. Wolbrink DRJ, van de Poll MCG, Termorshuizen F, de Keizer NF, van der Horst ICC, Schnabel R, Dejong CHC, van Santvoort HC, Besselink MG, van Goor H, et al. Trends in early and late mortality in patients with severe acute pancreatitis admitted to ICUs: A nationwide cohort study. Crit Care Med. 2022;50(10):1513–21.
2. Spanier B, Bruno MJ, Dijkgraaf MG. Incidence and mortality of acute and chronic pancreatitis in the Netherlands: a nationwide record-linked cohort study for the years 1995–2005. World J Gastroenterol. 2013;19(20):3018–26.
3. Schepers NJ, Bakker OJ, Besselink MG, Ahmed Ali U, Bollen TL, Gooszen HG, van Santvoort HC, Bruno MJ. Dutch pancreatitis study G: impact of characteristics of organ failure and infected necrosis on mortality in necrotising pancreatitis. Gut. 2019;68(6):1044–51.
4. Barreto SG, Kaambwa B, Venkatesh K, Sasson SC, Andersen C, Delaney A, Bihari S, Pilcher D, Collaborative PA. Mortality and costs related to severe acute pancreatitis in the intensive care units of Australia and new Zealand (ANZ), 2003–20. *Pancreatology* 2023.
5. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284(6):603–19.
6. Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? PLoS Med. 2018;15(12):e1002721.
7. Lilja HE, Leppaniemi A, Kemppainen E. Utilization of intensive care unit resources in severe acute pancreatitis. JOP. 2008;9(2):179–84.
8. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
9. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1.
10. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Medicine: Official J Soc Acad Emerg Med. 2011;18(10):1099–104.
11. Sun L, Yang H, Li J, Wang T, Li W, Liu G, Tang Y. In Silico prediction of compounds binding to human plasma proteins by QSAR models. ChemMedChem. 2018;13(6):572–81.
12. Chern CC, Chen YJ, Hsiao B. Decision tree-based classifier in providing telehealth service. BMC Med Inf Decis Mak. 2019;19(1):104.
13. Kaminska JA. A random forest partition model for predicting NO(2) concentrations from traffic flow and meteorological conditions. Sci Total Environ. 2019;651(Pt 1):475–83.
14. Bhosale H, Ramakrishnan V, Jayaraman VK. Support vector machine-based prediction of pore-forming toxins (PFT) using distributed representation of reduced alphabets. J Bioinform Comput Biol. 2021;19(5):2150028.
15. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care (London England). 2019;23(1):112.
16. Ochs RA, Goldin JG, Abtin F, Kim HJ, Brown K, Batra P, Roback D, McNitt-Gray MF, Brown MS. Automated classification of lung bronchovascular anatomy in CT using adaboost. Med Image Anal. 2007;11(3):315–24.
17. Su D, Zhang X, He K, Chen Y, Wu N. Individualized prediction of chronic kidney disease for the elderly in longevity areas in China: machine learning approaches. Front Public Health. 2022;10:998549.
18. Pereira F, Xiao K, Latino DA, Wu C, Zhang Q, Aires-de-Sousa J. Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. J Chem Inf Model. 2017;57(1):11–21.
19. Li K, Yao S, Zhang Z, Cao B, Wilson CM, Kalos D, Kuan PF, Zhu R, Wang X. Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics. 2022;38(6):1631–8.
20. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7(1):94.
21. Rodriguez-Perez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J Med Chem. 2020;63(16):8761–77.
22. Rahmatinejad Z, Peiravi S, Hoseini B, Rahmatinejad F, Eslami S, Abu-Hanna A, Reihani H. Comparing In-Hospital mortality prediction by senior emergency resident's judgment and prognostic models in the emergency department. Biomed Res Int. 2023;2023:6042762.
23. Li T, Huang H, Zhang S, Zhang Y, Jing H, Sun T, Zhang X, Lu L, Zhang M. Predictive models based on machine learning for bone metastasis in patients with diagnosed colorectal cancer. Front Public Health. 2022;10:984750.
24. Delpino FM, Costa AK, Farias SR, Chiavegatto Filho ADP, Arcencio RA, Nunes BP. Machine learning for predicting chronic diseases: a systematic review. Public Health. 2022;205:14–25.
25. Li X, Yang L, Yuan Z, Lou J, Fan Y, Shi A, Huang J, Zhao M, Wu Y. Multi-institutional development and external validation of machine learning-based models to predict relapse risk of pancreatic ductal adenocarcinoma after radical resection. J Translational Med. 2021;19(1):281.
26. Rahmatinejad Z, Hoseini B, Rahmatinejad F, Abu-Hanna A, Bergquist R, Pourmand A, Miri M, Eslami S. Internal validation of the predictive performance of models based on three ED and ICU scoring systems to predict inhospital mortality for intensive care patients referred from the emergency department. Biomed Res Int. 2022;2022:3964063.
27. Rahmatinejad Z, Reihani H, Tohidinezhad F, Rahmatinejad F, Peyravi S, Pourmand A, Abu-Hanna A, Eslami S. Predictive performance of the SOFA and mSOFA scoring systems for predicting in-hospital mortality in the emergency department. Am J Emerg Med. 2019;37(7):1237–41.
28. Rahmatinejad Z, Dehghani T, Hoseini B, Rahmatinejad F, Lotfata A, Reihani H, Eslami S. A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department. Sci Rep. 2024;14(1):3406.
29. Zhou Y, Ge YT, Shi XL, Wu KY, Chen WW, Ding YB, Xiao WM, Wang D, Lu GT, Hu LH. Machine learning predictive models for acute pancreatitis: A systematic review. Int J Med Inf. 2022;157:104641.
30. 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.
31. Wu BU, Johannes RS, Sun X, Conwell DL, Banks PA. Early changes in blood Urea nitrogen predict mortality in acute pancreatitis. Gastroenterology. 2009;137(1):129–35.
32. Al Mofleh IA. Severe acute pancreatitis: pathogenetic aspects and prognostic factors. World J Gastroenterol. 2008;14(5):675–84.
33. Wang Q, Chen Y, Huang P, Su D, Gao F, Fu X, Fu B. The clinical characteristics and outcome of elderly patients with acute pancreatitis. Pancreas. 2022;51(10):1284–91.
34. Kothari D, Struyvenberg MR, Perillo MC, Ezaz G, Freedman SD, Sheth SG. Extra-pancreatic complications, especially Hemodialysis predict mortality and length of stay, in ICU patients admitted with acute pancreatitis. Gastroenterol Rep (Oxf). 2018;6(3):202–9.
35. Kim YJ, Kim DB, Chung WC, Lee JM, Youn GJ, Jung YD, Choi S, Oh JH. Analysis of factors influencing survival in patients with severe acute pancreatitis. Scand J Gastroenterol. 2017;52(8):904–8.
36. Yuan H, Yu KY, Xie F, Liu MX, Sun SH. Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare. Med Adv. 2024;2(3):205–37.

Publisher's note