European Journal of Medical Research

RESEARCH

Open Access

Thirty-day mortality risk prediction for geriatric patients undergoing non-cardiac surgery in the surgical intensive care unit

Mengke Ma[1†], Jiatong Liu[2†], Caiyun Li[3], Yingxue Chen[4], Huishu Jia[1], Aijie Hou[3] and Hongzeng Xu[1,3*]

Abstract

Background The prediction of mortality for elderly patients undergoing non-cardiac surgeries is a vital research area, as accurate risk assessment can help surgeons make better clinical decisions during the perioperative period. This study aims to build a mortality risk prediction model for surgical intensive care unit (ICU) patients aged 65 and older undergoing non-cardiac surgery.

Methods Data was obtained from 1960 patients who underwent non-cardiac surgery from the medical information mart for intensive care IV (MIMIC-IV) database. The least absolute shrinkage selection operator (LASSO) regularization algorithm and the extreme gradient boosting (XGBoost) for feature importance evaluation were used to screen important predictors. Five predictive models were established: categorical boosting (CatBoost), logistic regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM). External validation was performed utilizing data from 153 patients in the MIMIC-III database. Finally, shapley additive explanations (SHAP) was utilized for a personalized analysis of the models.

Results Among the five predictive models developed in this study, the CatBoost model demonstrated superior overall performance in both the test data set (AUC = 0.96, F1 = 0.90) and the external validation data set (AUC = 0.98, F1 = 0.91). The decision curve analysis showed that the model offers a beneficial net benefit. The CatBoost model showed significant enhancements in classification accuracy when compared to the conventional revised cardiac risk index (RCRI) score. SHAP analysis revealed that anion gap, age, prothrombin time (PT), and weight were the four key variables influencing the predictive performance of the CatBoost model.

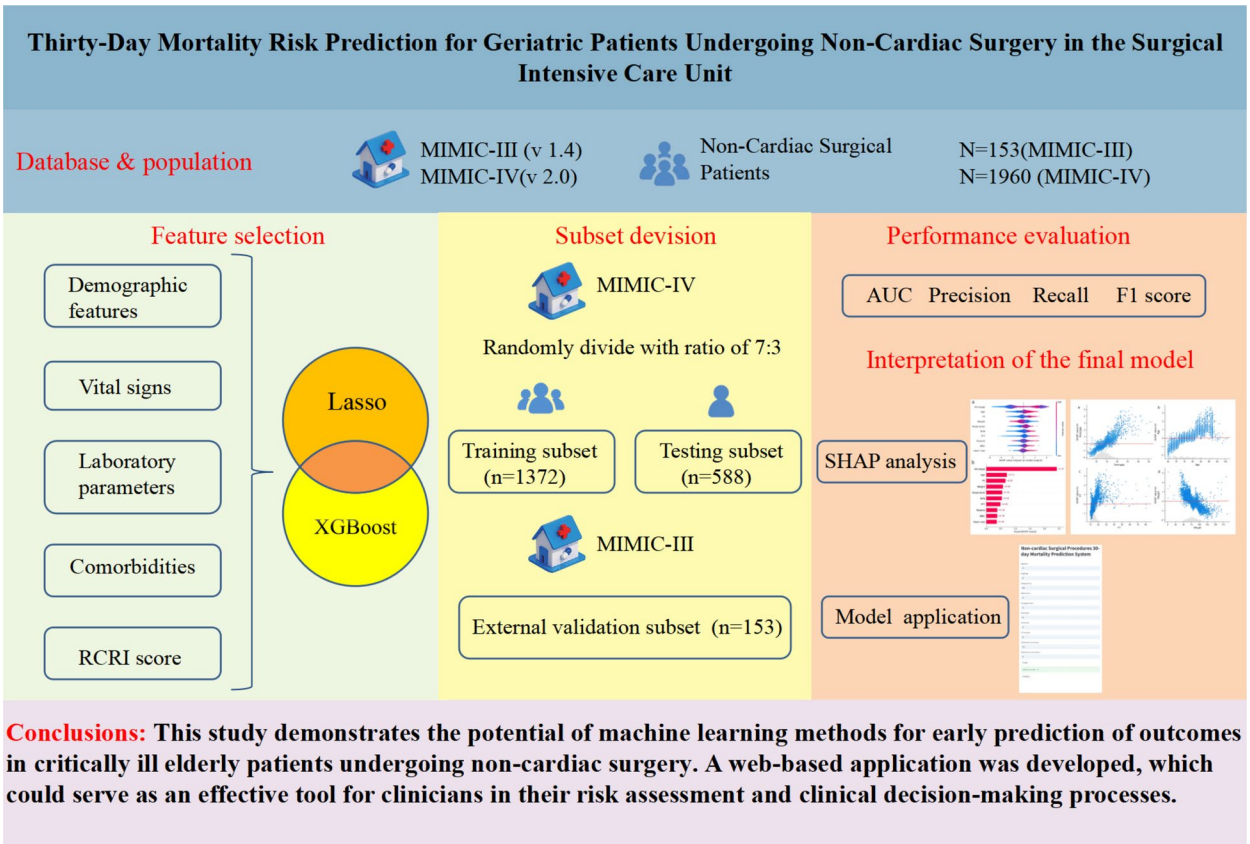
Conclusions This study demonstrates the potential of machine learning methods for early prediction of outcomes in critically ill elderly patients undergoing non-cardiac surgery. A web-based application was developed, which could serve as an effective tool for clinicians in their risk assessment and clinical decision-making processes.

Keywords Non-cardiac surgery, Machine learning, 30-Day mortality, MIMIC database

†Mengke Ma and Jiatong Liu have contributed equally to this work.

*Correspondence:
Hongzeng Xu
hongzengxu@foxmail.com
Full list of author information is available at the end of the article

Graphical abstract



Introduction

The global population is aging at an accelerating rate, with projections indicating that the proportion of individuals aged 65 and older will double from 10% in 2022 to 16% by 2050 [1]. This demographic shift underscores the escalating demand for healthcare interventions, particularly in the surgical domain. Consequently, the incidence of non-cardiac surgeries is soaring, accounting for approximately 85% of the total surgical volume [2]. A recent 7-day cohort study revealed that up to 8% of patients undergoing non-cardiac surgery (NCS) require critical care admission, with in-hospital mortality ranging from 1.2 to 21.5% [3]. These statistics highlight the significant risk associated with non-cardiac surgeries in elderly patients, necessitating robust preoperative risk assessment tools to optimize clinical decision-making and improve patient outcomes.

Currently, several tools are utilized to assess cardiac risk before surgery, including the revised cardiac risk index (RCRI) [4], the American College of Surgeons National Surgical Quality Improvement Program's myocardial infarction or cardiac arrest risk calculator (ACS–NSQIP–MICA) [5], and the ACS–NSQIP surgical risk calculator [6]. However, these tools are predominantly designed for cardiac surgeries and are not directly applicable to non-cardiac procedures. Moreover, traditional risk assessment methods, which often rely on expert experience and clinical trial results, are frequently hampered by high subjectivity and inconsistent outcomes. Existing predictive tools for non-cardiac surgeries exhibit a narrow range of accuracy in forecasting 30-day mortality outcomes [7]. For instance, the Surgical Outcome Risk Tools (SORT) lack external validation, the Apgar score for surgery is easy to apply but performs poorly in prediction, and the Physiological and Surgical Stress Score for Mortality and Morbidity (POSSUM) is burdensome due to data complexity and manual data handling requirements. These limitations underscore the urgent need for dedicated clinical risk assessment studies tailored to non-cardiac surgical

patients, aiming to bridge existing gaps and enhance the comprehensiveness and predictive efficiency of risk assessments.

Given these challenges, there is a compelling need for innovative approaches to improve preoperative risk prediction for elderly patients undergoing non-cardiac surgeries. Machine learning (ML) techniques offer a promising solution by leveraging large data sets to train models that can identify complex patterns and provide more accurate predictions. ML algorithms, such as random forest (RF), decision tree (DT), support vector machine (SVM), and neural networks (NN), have been extensively used in evaluating clinical risks. However, their application in predicting mortality outcomes specifically for elderly patients undergoing non-cardiac surgeries remains underexplored. Therefore, the present study aims to develop a machine learning-based model to forecast the risk of death in elderly patients after non-cardiac surgery, using preoperative medical history and various inspection indicators. This research is of significant relevance to clinical practice, academic research, and improving patient safety.

Methods

Database

The data used in this study was from the MIMIC database**.** From 2001 to 2012, the MIMIC-III database covered over 40,000 intensive care unit (ICU) patient cases at Beth Israel Deaconess Medical Center (BIDMC) [8]. Medical Information Mart for intensive care IV (MIMIC-IV) expanded on MIMIC-III and updated patient records from 2008 to 2019 [9]. Since the MIMIC series databases provide de-identified, publicly available data, this study did not require specific ethical approval [10]. In addition, to prevent the overlap of years, data from 153 patients in the MIMIC-III database spanning from 2001 to 2007 was selected for external validation.

Participants

The 2022 ESC/ESA non-cardiac surgical guidelines classify non-cardiac surgical procedures into low-risk, moderate-risk, and high-risk categories**.** Moderate-risk surgeries (e.g., peripheral arterial angioplasty, intraperitoneal, endovascular aneurysm repair, carotid endarterectomy, head and neck, gynecological, neurosurgical/major orthopedic and major urological) are those with intermediate risk of complications, while high-risk surgeries (e.g., adrenalectomy, aortic or major vascular resection, pancreaticoduodenectomy, hepatectomy, biliary, esophagectomy, pneumonectomy, lung transplantation, total cystectomy, and intestinal perforation repair) carry higher risk of morbidity and mortality [11]. This study focused on elective surgical patients admitted to the ICU while excluding other potential confounding factors and emergency surgery cases. The flowchart of the inclusion and exclusion criteria for this study is shown in Fig. 1. The exclusion criteria were: (1) only first-time ICU admissions included; (2) Excludes patients with ICU stays < 24 h [12]; (3) data from patients under 65 were ignored; (4) data from cardiac and other low-risk surgeries excluded; and (5) duplicate data avoided. Ultimately, 1960 patients met the criteria and were included, and the MIMIC-III data set followed the same inclusion and exclusion criteria.

Variable selection

In this study, we utilized software tools including PostgreSQL 13, pgAdmin 4 [13], and Python 3.9.7 to manage and organize data for variable selection. We executed SQL queries to extract the following clinical information:

1. Lab results on ICU Admission Day: hematocrit, hemoglobin, platelet count, red blood cells, white blood cells, anion gap, bicarbonate, creatinine, blood–urea–nitrogen, calcium, chloride, sodium, potassium, international normalized ratio (INR), prothrombin time (PT), first-day RCRI score, and activated partial thromboplastin time (APTT).
2. Demographics demographic information: Including patient age, sex, alcohol consumption, weight, height and BMI.
3. Vital signs averages on ICU Admission Day: body temperature, respiratory rate, systolic blood pressure, diastolic blood pressure, heart rate, oxygen saturation, and blood glucose levels.
4. Comorbidities: diabetes, liver disease, congestive heart failure, myocardial infarction, hypertension, renal disease, chronic lung disease, and cerebrovascular disease.

These data characteristics were essential for creating and confirming the accuracy of the forecasting model.

Data preprocessing and model construction

Outliers were directly excluded from the data set, and missing data issues in predictive variables were addressed using multiple imputation techniques [14]. Variables with a missing rate exceeding 30% were excluded to prevent bias and analysis complexity. Two feature selection algorithms were employed in this study, namely, least absolute shrinkage and selection operator (LASSO) regularization and extreme gradient boosting (XGBoost) feature importance ranking. These methods were utilized to identify key features and enhance model interpretability while preventing overfitting. In LASSO

regression, feature selection was conducted using the one-standard-error rule based on binary deviance. To enhance the interpretability of the XGBoost model, a threshold of 0.05 was set for feature importance, resulting in the selection of 15 significant features, with the final intersection of both algorithms identifying 10 factors significantly associated with 30-day mortality after non-cardiac surgery. These factors include anion gap, age, PT, weight, respiratory rate, blood–urea–nitrogen (BUN), APTT, platelet count, white blood cell count, and heart rate.

Given the significant disparity between the number of surviving patients and those who died postoperatively, resulting in imbalanced data, the synthetic minority over-sampling technique (SMOTE) was employed. This method increased the proportion of in-hospital death events relative to non-death events, adjusting the ratio from 1:12.9 to 1:1. The final data set used for experiments was split into training and testing sets in a 7:3 ratio, with the training set undergoing tenfold cross-validation [15, 16]. This method divided the training data into 10 equal parts, training the model on 9 parts and validating it on the remaining part, providing 10 independent performance estimates to help accurately assess the model's predictive capacity [17]. GridSearchCV was employed to meticulously tune hyperparameters for various machine learning models [18]. Through this process, the optimal parameter combinations were determined, and the evaluation metrics from GridSearchCV were recorded, which are presented in Table 1.

After the data preprocessing and feature selection, we trained five different models: logistic regression, decision tree, random forest, support vector machine (SVM), and CatBoost. We recorded the training and validation loss for each model during the training process and plotted separate loss curves for each model to monitor their convergence and performance. Each model was rigorously trained to ensure optimal performance and to provide a comprehensive comparison of their predictive capabilities.

Here are the formulas for five machine learning models [19]:

A. Logistic regression

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

This formula represents the probability that the logistic regression model predicts Y = 1. The logistic function maps the linear combination of features to a value between 0 and 1, representing the probability. Here, Y is the binary outcome variable, which takes the value 1 or 0, representing the two classes in a classification problem. $X1$, $X2$,..., Xn are the input features or predictor variables, and $\beta 0$, $\beta 1$,..., βn are the model parameters (coefficients)

Table 1 GridSearchCV optimization for machine learning models

ML models	Ranging value of GridSearchCV	Optimal parameter combination
LR	'C': [0.01, 0.1, 1, 10, 100]	'C': 0.1
	'penalty': ['l1','l2']	'penalty': l1
	'solver': ['liblinear','saga']	'solver': liblinear
	'n_estimators': [10, 50, 100, 200]	'n_estimators': 100
RF	'n_estimators': [10, 50, 100, 200]	'n_estimators': 100
	'max_depth': [5, 10, 15]	'max_depth': 10
	'min_samples_split': [2, 5, 10]	'min_samples_split': 2
	'min_samples_leaf': [1, 2, 4]	'min_samples_leaf': 1
DT DT	'max_depth': [3, 5, 10]	'max_depth': 10
	'min_samples_split': [2, 5, 10]	'min_samples_split': 5
	'min_samples_leaf': [1, 2, 4]	'min_samples_leaf': 2
SVM	'C': [0.1, 1, 10, 100]	'C': 10
	'kernel': ['linear','rbf','poly']	'kernel': rbf
	'gamma': ['scale', 'auto']	'gamma': auto
	'max_iter': [1000, 2000]	'max_iter': 1000
CatBoost CatBoost	'iterations': [10, 50, 100, 200]	'iterations': 200
	'depth': [5, 10, 15]	'depth': 10
	'learning_rate': [0.01, 0.1, 0.2]	'learning_rate': 0.1
	'l2_leaf_reg': [1, 3, 5, 7]	'l2_leaf_reg': 1

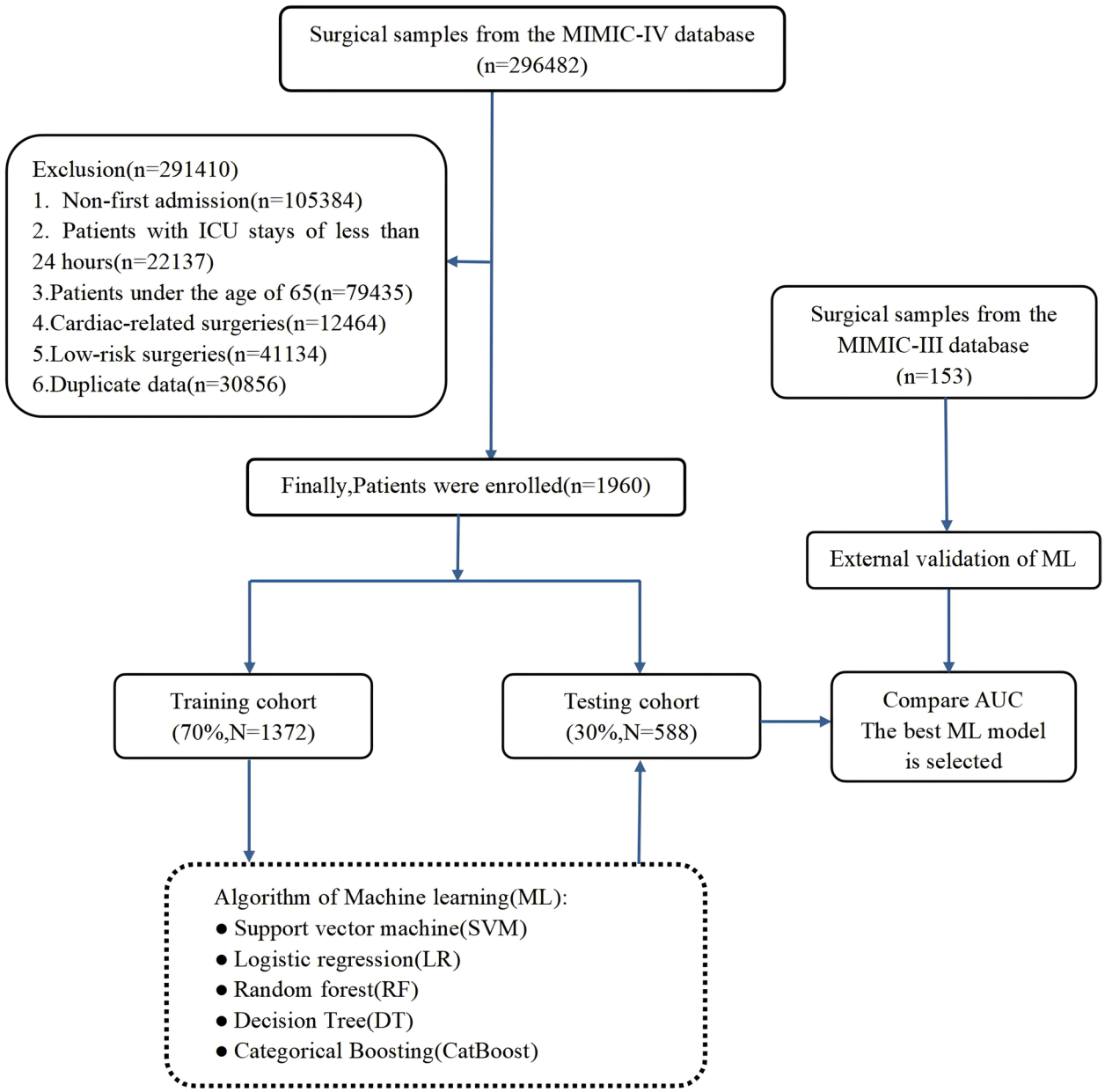


Fig. 1 Flowchart of the study

that determine the relationship between the input features and the probability of Y being 1. The term $e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}$ is the exponentiated linear combination of the input features, which is passed through the sigmoid function $\frac{1}{1+e^{-z}}$ to produce a probability value between 0 and 1.

B. Random forest

$$\hat{Y} = \text{mode}\{T_1(X), T_2(X), \ldots, T_k(X)\} \quad (2)$$

This formula represents the prediction result of the Random Forest model, where $\hat{Y}$ is the predicted output for a given input X, which is determined by majority voting. In this formula, T_1 (X), T_2 (X), ..., T_k (X) are the predictions made by each individual decision tree T_1, T_2,..., T_k in the forest, based on the same input features X. The mode is the mode function, which returns the most frequent predicted class.

C. CatBoost

$$F(x) = \sum_{t=1}^{T} \eta_t h_t(x) \quad (3)$$

This formula represents the prediction result of the CatBoost model, where $F(x)$ is the final predicted value, T is the number of trees in the model, $h_t(x)$ is the prediction of the tth tree, η_t is the learning rate, which controls the contribution of each tree.

D. Support vector machine (SVM)

$$f(x) = w^T x + b \tag{4}$$

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \tag{5}$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{6}$$

This formula (4) represents the decision function of the SVM model, where w is the weight vector, x is the feature vector, and b is the bias term. The sign of the function determines the classification result: when $f(x) > 0$, it means the sample belongs to class + 1; When $f(x) < 0$, it means the sample belongs to class − 1. The formula (5) represents the objective function of the SVM optimization problem, where $\| w \|^2$ is the squared norm of the weight vector w, representing the margin between classes, and C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing classification errors. The summation $\sum_{i=1}^{n} \xi_i$ accounts for the slack variables ξ_i, which allow some misclassification or margin violation. The formula (6) represents the constraints in the SVM optimization, where y_i is the class label of the ith sample, x_i is the feature vector, w is the weight vector, and b is the bias term. This ensures that correctly classified samples are on the correct side of the margin, with a minimum distance of 1, and the slack variables ξ_i are non-negative, allowing for some margin violations.

E. Decision tree

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{7}$$

This formula represents the prediction result of the Decision Tree model, which returns the constant value ci based on the region Ri that the input X falls into, where $IG(D_p, f)$ represents the information gain when splitting the data set D_p based on feature f. In the process of building a decision tree, information gain is used to select the best feature for data partitioning. A larger information gain indicates that the feature can more effectively separate the data, making the tree-building process more efficient and thereby improving the model's predictive ability. $I(D_p)$ is the entropy of the parent node D_p, which is used to evaluate the splitting effect between the parent node and the child nodes, helping to select the splitting point. N_p is the total number of samples in the parent node. The summation $\sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$ calculates the weighted sum of the entropies of the child nodes D_j, where N_j is the number of samples in the jth child node and $I(D_j)$ is the entropy of that node. Through weighted calculation, the decision tree can better evaluate the effectiveness of each splitting step, thus achieving optimal partitioning.

Model evaluation

The evaluation metrics include accuracy, recall, F1 score, and the area under the receiver operating characteristic curve (AUC), with AUC and F1 score being the primary indicators for assessing model performance. To improve the evaluation of the model's utility and accuracy in a clinical setting, clinical decision curves and calibration curves were constructed for analysis. In these indicators, TP represents the True Positive, TN represents the True Negative, FP represents the False Positive, and FN represents the False Negative. The formulas for these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}. \tag{11}$$

Statistical analysis

This study indicates that the collected data follow a normal distribution pattern. Categorical variables are expressed as percentages, while continuous variables are described as mean ± standard deviation [20]. The Kolmogorov–Smirnov test was used for continuous variables, and the chi-square test was applied to categorical variables. All statistical analyses were performed using Python [21], and a p value < 0.05 was considered statistically significant.

Results

Baseline characteristics

After screening 296,482 samples from the MIMIC-IV database, the study ultimately included 1960 participants. As shown in Table 2, the average age of the subjects was 76.33 ± 7.60 years, with an average survival age of 75.78 ± 7.42 years and an average age at death of

Table 2 Baseline characteristics of the patients

	Total *N* = 1960	Non-survivors *N* = 224	Survivors *N* = 1736	*p* value
Gender				0.006
Female	1040 (53.1)	99 (44.2)	941 (54.2)	
Male	920 (46.9)	125 (55.8)	795 (45.8)	
Age, years	76.33 (7.60)	80.56 (7.72)	75.78 (7.42)	< 0.001
Alcohol				0.846
Never	1753 (89.4%)	199 (88.8%)	1554 (89.5%)	
Ofen	207 (10.6%)	25 (11.2%)	182 (10.5%)	
Height (cm)	167.97 (11.06)	166.44 (11.29)	168.17 (11.01)	0.028
Weight (kg)	79.93 (19.99)	76.09 (22.38)	80.43 (19.61)	0.002
BMI, kg/m^2	28.46 (7.23)	27.61 (8.20)	28.57 (7.09)	0.061
First day RCRI score	1.60 (1.22)	2.75 (1.22)	1.45 (1.14)	< 0.001
Hematocrit (%)	31.39 (5.06)	30.88 (5.33)	31.45 (5.02)	0.109
Hemoglobin (g/dL)	10.40 (1.73)	10.09 (1.79)	10.43 (1.72)	0.006
Platelet ($\times 10^9$/L)	202.07 (98.01)	219.76 (117.57)	199.78 (94.99)	0.004
Red blood cells ($\times 10^{12}$/L)	3.33 (0.52)	3.22 (0.54)	3.34 (0.51)	0.001
White blood cells ($\times 10^9$/L)	13.35 (10.54)	15.77 (11.53)	13.04 (10.36)	< 0.001
Anion gap (mmol/L)	13.87 (3.39)	16.02 (4.03)	13.59 (3.20)	< 0.001
Bicarbonate (mmol/L)	22.68 (3.38)	21.37 (4.15)	22.85 (3.23)	< 0.001
BUN (mg/dL)	23.47 (15.68)	32.32 (21.42)	22.33 (14.39)	< 0.001
Calcium (mg/dL)	8.24 (0.66)	8.17 (0.74)	8.25 (0.64)	0.065
Chloride (mg/dL)	104.94 (4.75)	104.29 (6.14)	105.03 (4.54)	0.029
Creatinine (mg/dL)	1.26 (1.08)	1.59 (1.21)	1.22 (1.05)	< 0.001
Sodium (mmol/L)	138.42 (3.89)	138.57 (4.74)	138.40 (3.77)	0.526
Potassium (mmol/L)	4.26 (0.54)	4.35 (0.68)	4.25 (0.52)	0.009
INR	1.38 (0.44)	1.60 (0.66)	1.35 (0.39)	< 0.001
PT	14.97 (4.59)	17.32 (6.89)	14.67 (4.11)	< 0.001
APTT	35.73 (16.02)	40.03 (18.85)	35.17 (15.54)	< 0.001
Heart rate	83.09 (13.99)	86.86 (16.12)	82.60 (13.62)	< 0.001
Systolic blood pressure	118.46 (15.07)	117.03 (17.77)	118.64 (14.68)	0.131
Diastolic blood pressure	58.56 (9.28)	58.56 (9.87)	58.56 (9.20)	0.996
Respiration	18.45 (3.20)	19.83 (3.54)	18.27 (3.11)	< 0.001
Temperature	36.80 (0.43)	36.68 (0.57)	36.81 (0.41)	< 0.001
Pulse oxygen saturation	97.20 (1.88)	96.96 (2.76)	97.23 (1.73)	0.047
Glucose (mmol/L)	14.08 (2.21)	14.84 (2.87)	8.18 (2.34)	0.64
Myocardial infarct				0.22
No	1611 (82.2%)	177 (79.0%)	1434 (82.6%)	
Yes	349 (17.8%)	47 (21.0%)	302 (17.4%)	
Congestive heart failure				< 0.001
No	1462 (74.6%)	137 (61.2%)	1325 (76.3%)	
Yes	498 (25.4%)	87 (38.8%)	411 (23.7%)	
Hypertension				0.034
No	941 (48.0%)	123 (54.9%)	818 (47.1%)	
Yes	1019 (52.0%)	101 (45.1%)	918 (52.9%)	
Cerebrovascular disease				0.002
No	1604 (81.8%)	166 (74.1%)	1438 (82.8%)	
Yes	356 (18.2%)	58 (25.9%)	298 (17.2%)	
Chronic pulmonary disease				0.105
No	1440 (73.5%)	154 (68.8%)	1286 (74.1%)	

Table 2 (continued)

	Total N = 1960	Non-survivors N = 224	Survivors N = 1736	p value
Yes	520 (26.5%)	70 (31.2%)	450 (25.9%)	
Diabetes				0.499
No	1375 (70.2%)	162 (72.3%)	1213 (69.9%)	
Yes	585 (29.8%)	62 (27.7%)	523 (30.1%)	
Renal disease				< 0.001
No	1562 (79.7%)	155 (69.2%)	1407 (81.0%)	
Yes	398 (20.3%)	69 (30.8%)	329 (19.0%)	
Liver disease				< 0.001
No	1795 (91.6%)	188 (83.9%)	1607 (92.6%)	
Yes	165 (8.4%)	36 (16.1%)	129 (7.4%)	

Table 3 Performances of the five machine learning models and the RCRI score

Model	AUC	Precision	Recall	F1 score
CatBoost	0.96	0.86	0.94	0.90
RCRI score	0.78	0.68	0.67	0.68
Logistic regression	0.78	0.70	0.68	0.70
Random forest	0.94	0.87	0.86	0.86
SVM	0.91	0.80	0.88	0.84
Decision tree	0.84	0.77	0.78	0.77

80.56 ± 7.72 years. The gender distribution was 53.1% female (1040 participants) and 46.9% male (920 participants). Hypertension was the most common condition, affecting 1019 participants (52.0% of the group). Among the selected non-cardiac surgery patients, 224 died, accounting for 11.4%. Significant statistical differences ($p < 0.001$) were found in variables, such as age, white blood cell count, first day RCRI score, anion gap, blood–urea–nitrogen, INR, and PT.

Model performance

The results (Table 3) show that the CatBoost model achieved the highest performance in the test set (AUC = 0.96, F1 = 0.90), followed by the random forest model (AUC = 0.94, F1 = 0.86), the SVM model (AUC = 0.91, F1 = 0.84), and the decision tree model (AUC = 0.84, F1 = 0.77). These results highlight the superior predictive accuracy and efficiency of the CatBoost model in this context. The MIMIC-III database was used to further validate performance advantage of CatBoost. Figure 2b shows that CatBoost continued to exhibit a high AUC (0.98), outperforming other models (RF, 0.97; LR, 0.78; SVM, 0.96; DT, 0.82), confirming its excellent classification performance.

The training and validation loss curves for the five models—Logistic Regression, Decision Tree, Random Forest, SVM, and CatBoost—are presented in Additional file 4. All models showed a decrease in both training and validation loss as the number of training samples increased, indicating convergence. However, the training loss consistently decreased faster and reached lower values compared to the validation loss, suggesting some degree of overfitting. CatBoost demonstrated the lowest validation loss, indicating the best generalization performance among the models.

In addition, logistic regression was utilized with RCRI scores as a continuous variable to calculate the AUC for the Receiver Operating Characteristic curve [22], which was only 0.78 (Additional file 2). This further confirms that the CatBoost model surpasses the traditional RCRI score in both accuracy and predictive capability.

Due to its exceptional performance, CatBoost model was selected for further predictive analysis. The calibration curve shown in Fig. 2c indicates that the CatBoost model had a calibration slope of 0.0443, lower than that of other models. Calibration shows the gap between predicted and actual probabilities, where a lower value signifies higher model precision. The CatBoost model consistently displayed better fit than other models, as shown in Fig. 2d.

Model interpretation

This study performed SHAP (SHapley Additive exPlanations) analyses on the 10 predictive variables to evaluate their individual impact on the CatBoost model's forecasts. SHAP values, which are based on game theory and represent the average contribution of a feature to the model's prediction across all possible feature combinations, were used to assess the influence of each feature. Figure 3a shows the SHAP values, where each feature's horizontal position on the plot signifies

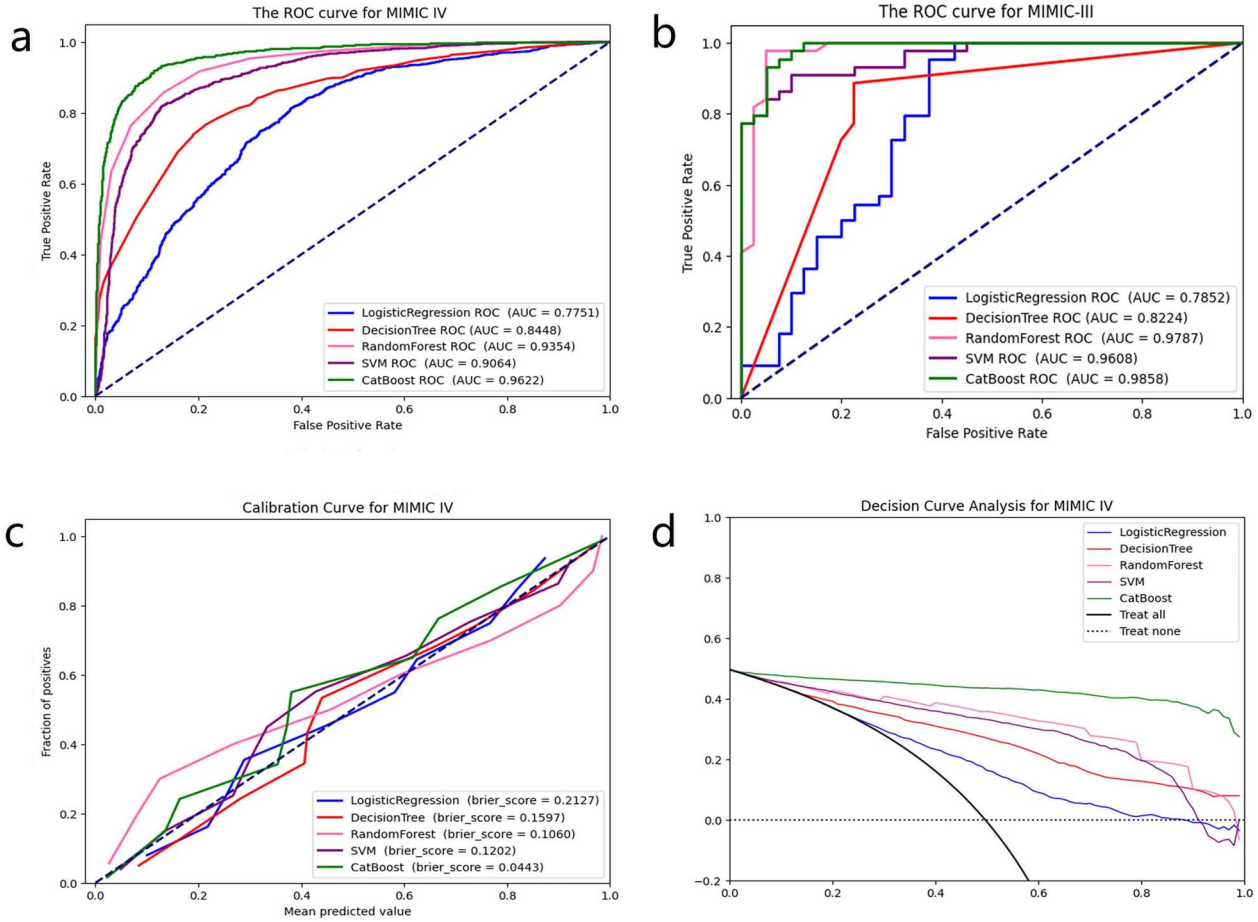


Fig. 2 Model performance evaluation. **a** ROC curve for test cohort. **b** ROC curve for the external validation cohort. **c** Calibration curve for the test cohort. **d** Decision curve analysis (DCA) for the test cohort

its association with increased or decreased predictive trends. Factors such as age, platelet count, white blood cell count, anion gap, blood–urea–nitrogen, PT, activated partial thromboplastin time (APTT), heart rate, and respiratory rate positively influence mortality risk predictions. Conversely, an increase in weight negatively impacts survival predictions.

In Fig. 3b, the bar chart lists features from highest to lowest based on their average absolute SHAP values. The top four features are Anion Gap, Age, PT, and Weight, with absolute SHAP values of 2.42, 0.71, 0.66, and 0.58, respectively. This ordering shows the contribution of every feature to the model's performance. A higher absolute SHAP value signifies greater importance and a stronger influence on the model's output.

Furthermore, SHAP dependence plots were generated for the top four influential clinical features to explain their impact on patient mortality risk [23]. Figure 4a shows the relationship between anion gap (horizontal axis) and mortality risk (vertical axis), revealing a critical value at 20 mmol/L. This indicates that patients with a value above this threshold have an increased risk of death within 30 days following non-cardiac surgery. Similarly, Fig. 4b–d indicates critical values for age (80 years, positively correlated), PT (20 s, positively correlated) and weight (75 kg, negatively correlated).

Model application

This study employed the CatBoost algorithm due to its superior AUROC performance on both the test and validation sets and developed an online web calculator specifically designed for clinical applications (http://39.99.156.31:10052/) [24]. As shown in Additional file 3, clinicians can input relevant patient data into the calculator to obtain predictive results, which aids

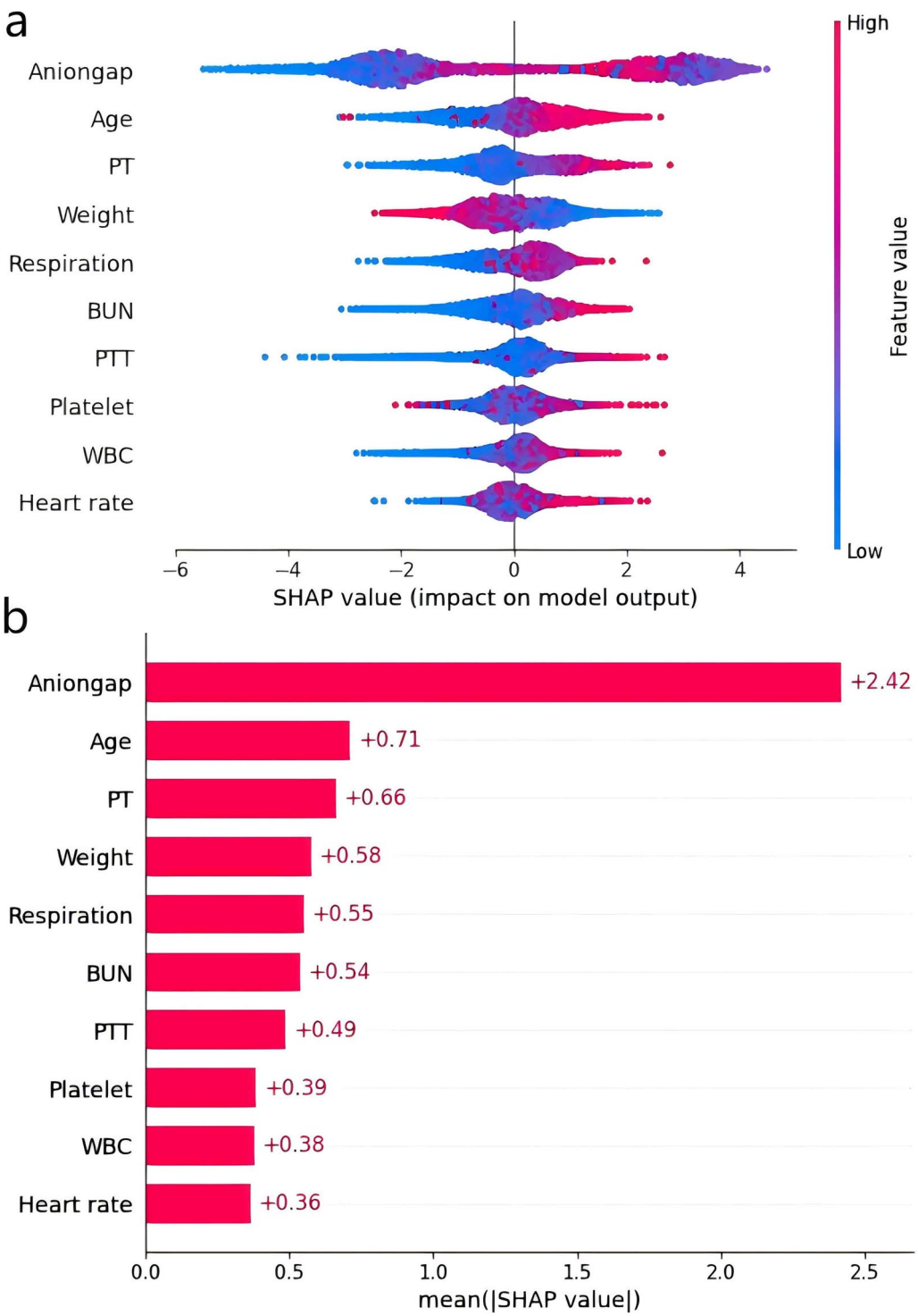


Fig. 3 Explaining model predictions with SHAP values and feature importance in CatBoost. **a** SHAP values of the eleven predictors. **b** Feature importance of the CatBoost model

clinical practitioners in making optimal decisions. This tool aims to support clinical decision-making, enhance perioperative decision processes, and optimize the allocation of healthcare resources.

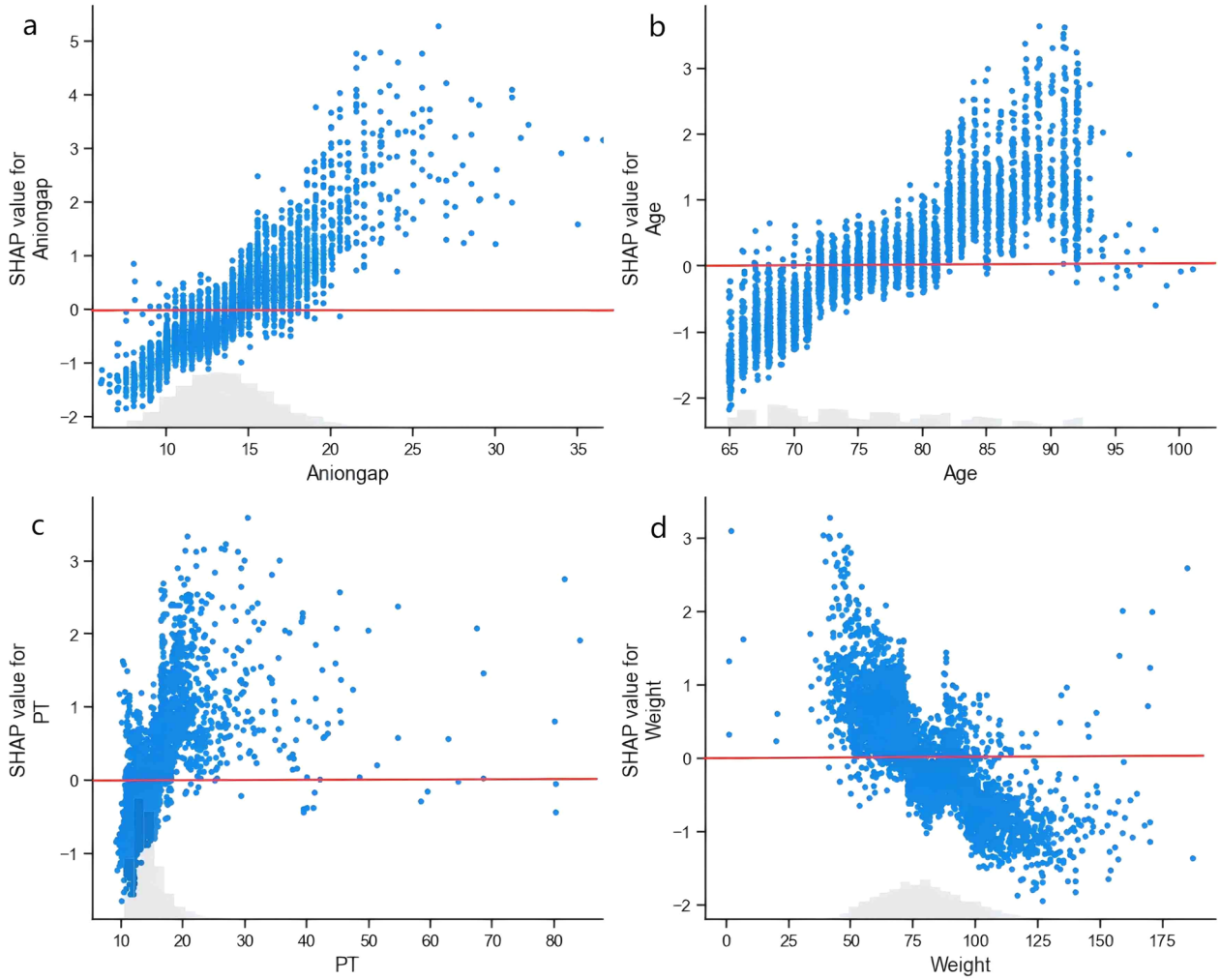


Fig. 4 SHAP dependency plots for the top 4 clinical features influencing model output. **a** SHAP Dependency Plot for Anion Gap. **b** SHAP Dependency Plot for Age. **c** SHAP Dependency Plot for PT. **d** SHAP Dependency Plot for Weight

Discussion

In this study, five machine-learning were developed and validated algorithms for predicting 30-day mortality following non-cardiac surgery using the MIMIC database. Among these models, CatBoost showed better results than SVM, RF, LR and DT. External validation using the MIMIC-III database further confirmed the superior performance of CatBoost.

The RCRI, thoroughly revised in 1999, includes six critical factors: preoperative serum creatinine level > 2 mg/dL, coronary artery disease, cerebrovascular events, insulin use, persistent heart failure, and high-risk surgical procedures, such as abdominal, thoracic, or groin vascular surgery [25]. Subsequent studies have validated this index, demonstrating its moderate predictive ability for cardiac mortality and nonfatal cardiac arrest in non-cardiac surgery patients [26]. This study found that the CatBoost model outperformed the traditional RCRI scoring system, highlighting its advantages in predictive accuracy. CatBoost, a resource-efficient and scalable alternative to learn models, offers optimal performance with relatively minimal computational resources and data requirements. It has become an important tool for predicting in-hospital mortality risk in critical care settings, assisting clinicians in decision-making.

The results (Table 4) show that the training times for logistic regression and decision tree were 0.2312 s and 0.0332 s, respectively. Their relatively low time complexity makes them suitable for large-scale data sets, but they may fail to capture complex nonlinear relationships. Despite its higher time complexity, random forest is widely used in practical applications due to its strong generalization ability and parallelizable characteristics.

SVM, with a training time of 4.3109 s, has relatively high time complexity, and the training time can be very long, especially when using nonlinear kernel functions. CatBoost, with a training time of 0.6999 s, has time complexity similar to that of decision trees and random forests. However, through algorithmic optimizations and hardware acceleration, its training time is relatively lower when dealing with categorical features and large-scale data sets [27]. Therefore, when selecting a model, it is necessary to consider factors, such as time complexity, model accuracy, data scale, and real-time requirements.

Based on the MIMIC database, the study identified 10 risk factors associated with a higher likelihood of 30-day mortality following non-cardiac surgery in elderly patients. Many of these factors are related to laboratory test results, underscoring the importance of preoperative lab tests [28, 29]**.** Addressing these factors prior to surgery can significantly decrease the likelihood of mortality during the perioperative period. Other studies have shown that high-sensitivity cardiac troponin *T*, total cholesterol, and high-density lipoprotein (HDL) levels are highly associated with 6-month mortality [30]. These confirmed risk factors provide a solid foundation for comprehensive preoperative assessment and provide the necessary information to help clinical doctors make scientific decisions. Among the identified risk factors, the most critical predictors were anion gap, age, PT and weight. The anion gap is a key marker for assessing metabolic acidosis and identifying potential pathogenic conditions. The association between elevated anion gap and increased mortality in hospitalized patients has been demonstrated [31]. Increasing age often correlates with multiple underlying conditions, raising the risk of postoperative mortality. Significantly higher complication rates and mortality have been reported among patients aged 80 and above compared to younger groups [32–34]. Similarly, Troisi determined a significant uptick in death risks for elderly individuals receiving intensive care [35]. Figure 4b shows a significantly higher likelihood of death among individuals over 80 years of age, further validating age's critical role in postoperative mortality prediction. PT, a crucial indicator for assessing coagulation function, was found to significantly impact mortality risk [36], with a PT value exceeding 20 s indicating a substantial increase in risk (Fig. 4c). Although heavier patients typically face higher surgical mortality due to associated comorbidities [37], this study found an unexpected result: higher weight did not raise the mortality risk, indicating a need for further study.

Accurate risk prediction is crucial for elderly patients undergoing non-cardiac surgery, given their increased vulnerability to postoperative complications and mortality due to multiple comorbidities. Effective risk prediction models can significantly enhance perioperative management and improve outcomes by facilitating targeted interventions and optimizing clinical decision-making. In the preoperative phase, identifying high-risk patients through risk prediction models allows for early, tailored interventions, such as optimizing treatment plans and adjusting medications, thereby reducing the incidence of perioperative complications and enhancing surgical safety. During the perioperative period, real-time risk prediction enables dynamic monitoring and early detection of potential risks, facilitating timely interventions and reducing postoperative mortality and hospital stay duration. In the postoperative phase, risk prediction models aid in closely monitoring high-risk patients, enabling prompt management of complications. In addition, these models provide psychological support to patients and their families, enhancing satisfaction and compliance.

In summary, comprehensive risk prediction and management strategies can significantly improve postoperative outcomes and quality of life for elderly patients undergoing non-cardiac surgery. Future work should focus on integrating advanced predictive analytics into clinical workflows to further enhance patient care.

While this study provides valuable insights into the 30-day mortality risk prediction for geriatric patients undergoing non-cardiac surgery, several limitations should be acknowledged. First, the data set included patients aged 65–89 years, which may limit the generalizability of the findings to other age groups. Future studies should consider a broader age range to enhance the applicability of the model. Second, data with missing values exceeding 30% were excluded, potentially introducing selection bias and leading to an incomplete data set. Although this approach helps reduce the impact of missing data on model performance, it may exclude valuable information. Third, the model was developed using data from the first ICU admission, lacking subsequent dynamic data. This might limit the model's ability to capture the full trajectory of a patient's condition over time. Incorporating longitudinal data could provide a more comprehensive view of the patient's status and potentially improve predictive accuracy. Fourth, the

data were sourced from the MIMIC database, specific to ICU populations, which may limit the universality of the findings. Future studies should consider multicenter data from diverse populations to validate the model's effectiveness in different settings [38]. Fifth, the model's implementation and visualization rely on Python software, which might present usability challenges for clinicians unfamiliar with this programming environment. Sixth, the sample size, although relatively large, may still be insufficient for certain subgroup analyses, potentially affecting the statistical power and reliability of the results. Future studies should aim for larger sample sizes to ensure robust and generalizable findings.

In conclusion, while this study demonstrates the potential of machine learning models for predicting 30-day mortality in geriatric patients undergoing non-cardiac surgery, addressing the aforementioned limitations in future research is essential to enhance the model's robustness and applicability.

Conclusion

This study successfully developed a machine learning-based model to predict the 30-day mortality rate of elderly patients undergoing non-cardiac surgery. The model demonstrated superior performance compared to traditional risk assessment tools and was further validated using an external data set. In addition, a web-based calculator was created to enhance the model's usability, allowing clinicians to assess patient risk before non-cardiac surgery. The primary purpose of this tool is to support clinical decision-making by providing accurate risk predictions, thereby optimizing the allocation of medical resources and improving patient outcomes. Future work should focus on validating the model in larger, multicenter cohorts to further assess its generalizability and clinical applicability.

Appendix

See Table 4.

Table 4 Training time of five models

Model	Training time (s)
CatBoost	0.6999
Logistic regression	0.2312
Random forest	0.1285
SVM	4.3109
Decision tree	0.0332

Abbreviations

ICU	Intensive care unit
LASSO	Least absolute shrinkage and selection operator
XGBoost	Extreme gradient boosting
DT	Decision tree
RF	Random forest
CatBoost	Categorical boosting
LR	Logistic regression
SHAP	Shapley additive explanations
RCRI	Revised cardiac risk index
PT	Prothrombin time
NN	Nural networks
INR	International normalized ratio
APTT	Activated partial thromboplastin time
HDL	High-density lipoprotein
AUC	Area under the curve
ROC	Receiver operating characteristic
SMOTE	Synthetic minority over-sampling technique
BIDMC	Beth Israel Deaconess Medical Center
MIMIC-IV	Medical information mart for intensive care IV
NSQIP	National surgical quality improvement program
ACS	American college of surgeons
LR	Logistic regression
SVM	Support vector machine
BUN	Blood–urea–nitrogen
NCS	Non-cardiac surgery

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-025-02543-1.

Additional file 1

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Author contributions

MKM contributed to the study conception and design. Material preparation, data collection and analysis were performed by JTL and HSJ. The draft of the manuscript was written by CYL and YXC. HZX and AJH revised the manuscript. All authors read and approved the final manuscript.

Funding

This work received funding from the Natural Science Foundation of Liaoning Province under Grant (No. 2023-MS-054).

Availability of data and materials

The data sets analyzed during the current study can be obtained from (https://physionet.org/content/mimiciii/1.4/) and (https://physionet.org/content/mimiciv/2.0/).

Declarations

Ethics approval and consent to participate

Mengke Ma has completed the online test and received a certification number, which is Record ID: 59607662. However, we adhered to fundamental ethical research principles to ensure the validity and fairness of the investigation.

Consent for publication

All authors have thoroughly reviewed the full manuscript and agreed to the publication of the final version.

Competing interests
The authors declare no competing interests.

Author details
[1]The People's Hospital of Liaoning Province, Command Postgraduate Training Base, Jinzhou Medical University, Shenyang, China. [2]Department of Anesthesiology, General Hospital of Northern Theater Command, 83 Wenhua Road, Shenyang 110016, Liaoning Province, China. [3]Department of Cardiology, The People's Hospital of Liaoning Province, 33 Wenyi Road, Shenyang 110011, Liaoning Province, China. [4]School of Intelligent Medicine, China Medical University, Puhe Road, Shenyang 110122, Liaoning Province, China.

Received: 2 January 2025 Accepted: 31 March 2025
Published online: 09 May 2025

References

1. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med. 2023;29:1623–30.
2. Halvorsen S, Mehilli J, Cassese S, Hall TS, Abdelhamid M, Barbato E, De Hert S, de Laval I, Geisler T, Hinterbuchner L, et al. 2022 ESC Guidelines on cardiovascular assessment and management of patients undergoing non-cardiac surgery. Eur Heart J. 2022;43:3826–924.
3. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, Vallet B, Vincent J-L, Hoeft A, Rhodes A. Mortality after surgery in Europe: a 7 day cohort study. The Lancet. 2012;380:1059–65.
4. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, Sugarbaker DJ, Donaldson MC, Poss R, Ho KK, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation. 1999;100:1043–9.
5. Davenport DL, Bowe EA, Henderson WG, Khuri SF, Mentzer RM. National surgical quality improvement program (NSQIP) risk factors can be used to validate american society of anesthesiologists physical status classification (ASA PS) levels. Ann Surg. 2006;243:636–44.
6. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, Cohen ME. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013;217:833-842e833.
7. Woo SH, Marhefka GD, Cowan SW, Ackermann L. Development and validation of a prediction model for stroke, cardiac, and mortality risk after non-cardiac surgery. J Am Heart Assoc. 2021;10: e018013.
8. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:1–9.
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.
10. Zhang G, Shao F, Yuan W, Wu J, Qi X, Gao J, Shao R, Tang Z, Wang T. Predicting sepsis in-hospital mortality with machine learning: a multicenter study using clinical and inflammatory biomarkers. Eur J Med Res. 2024;29:156.
11. Wu X, Hu J, Zhang J. Machine learning-based model for predicting major adverse cardiovascular and cerebrovascular events in patients aged 65 years and older undergoing noncardiac surgery. BMC Geriatr. 2023;23:819.
12. Liu S, Chen M, Tang L, Li X, Zhou S. Association between serum ferritin and prognosis in patients with ischemic heart disease in intensive care units. J Clin Med. 2023;12:6547.
13. Dou J, Guo C, Wang Y, Peng Z, Wu R, Li Q, Zhao H, Song S, Sun X, Wei J. Association between triglyceride glucose-body mass and one-year all-cause mortality of patients with heart failure: a retrospective study utilizing the MIMIC-IV database. Cardiovasc Diabetol. 2023;22:309.
14. Dhanka S, Maini S. A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. Int J Cardiol. 2025;420: 132757.
15. Allgaier J, Pryss R. Cross-validation visualized: a narrative guide to advanced methods. Mach Learn Knowl Extract. 2024;6:1378–88.
16. Dhanka S, Maini S. Random forest for heart disease detection: a classification approach. in book random forest for heart disease detection: a classification approach (Editor eds.); 2021. p. 1–3.
17. Huang J, Cai Y, Wu X, Huang X, Liu J, Hu D. Prediction of mortality events of patients with acute heart failure in intensive care unit based on deep neural network. Comput Methods Programs Biomed. 2024;256: 108403.
18. Dhanka S, Bhardwaj VK, Maini S. Comprehensive analysis of supervised algorithms for coronary artery heart disease detection. Expert Syst. 2023;40: e13300.
19. Sharma A, Dhanka S, Kumar A, Maini S. A comparative study of heterogeneous machine learning algorithms for arrhythmia classification using feature selection technique and multi-dimensional datasets. Eng Res Express. 2024;6: 035209.
20. Çiçek V, Cinar T, Hayiroglu MI, Kılıç Ş, Keser N, Uzun M, Orhan AL. Preoperative cardiac risk factors associated with in-hospital mortality in elderly patients without heart failure undergoing hip fracture surgery: a single-centre study. Postgrad Med J. 2021;97:701–5.
21. Peng X, Zhu T, Chen Q, Zhang Y, Zhou R, Li K, Hao X. A simple machine learning model for the prediction of acute kidney injury following noncardiac surgery in geriatric patients: a prospective cohort study. BMC Geriatr. 2024;24:549.
22. Palamuthusingam D, Pascoe EM, Hawley CM, Johnson DW, Fahim M. Revised cardiac risk index in predicting cardiovascular complications in patients receiving chronic kidney replacement therapy undergoing elective general surgery. Perioperative Medicine. 2024;13:70.
23. 龚欢欢, 柯晓伟, 王爱民, 李湘民: 可解释机器学习模型预测心脏骤停患者院内死亡风险. Med J Pek Union Med Coll Hosp. 2023;14:528–35.
24. Li M, Han S, Liang F, Hu C, Zhang B, Hou Q, Zhao S. Machine learning for predicting risk and prognosis of acute kidney disease in critically ill elderly patients during hospitalization: internet-based and interpretable model study. J Med Internet Res. 2024;26: e51354.
25. Schmidt G, Frieling N, Schneck E, Habicher M, Koch C, Aßmus B, Sander M. Comparison of preoperative NT-proBNP and simple cardiac risk scores for predicting postoperative morbidity after non-cardiac surgery with intermediate or high surgical risk. Perioper Med. 2024;13:44.
26. Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. Ann Intern Med. 2010;152:26–35.
27. Dhanka S, Maini S. HyOPTXGBoost and HyOPTRF: hybridized intelligent systems using optuna optimization framework for heart disease prediction with clinical interpretations. Multimed Tools Appl. 2024;83:72889–937.
28. Clerico A, Zaninotto M, Aimo A, Musetti V, Perrone M, Padoan A, Dittadi R, Sandri MT, Bernardini S, Sciacovelli L, et al. Evaluation of the cardiovascular risk in patients undergoing major non-cardiac surgery: role of cardiac-specific biomarkers. Clin Chem Lab Med (CCLM). 2022;60:1525–42.
29. Martin SK, Cifu AS. Routine preoperative laboratory tests for elective surgery. JAMA. 2017;318:567–8.
30. Wu XD, Wang Q, Song YX, Chen XY, Xue T, Ma LB, Luo YG, Li H, Lou JS, Liu YH, et al. Risk factors prediction of 6-month mortality after noncardiac surgery of older patients in China: a multicentre retrospective cohort study. Int J Surg (Lond). 2024;110:219–28.
31. Ceruti S, Yang DE, Jo S, Lee DH, An WS, Jeong MJ, Son M. Dynamics of serum anion gaps with in-hospital mortality: analysis of the multi-open databases. PLoS ONE. 2024;19:1.
32. Machado AN, do Carmo Sitta M, Jacob Filho W, Garcez-Leme LE. Prognostic factors for mortality among patients above the 6th decade undergoing non-cardiac surgery: (cares-clinical assessment and research in elderly surgical patients). Clin Sci. 2008;63:151–6.
33. Cicek V, Babaoglu M, Saylik F, Yavuz S, Mazlum AF, Genc MS, Altinisik H, Oguz M, Korucu BC, Hayiroglu MI, et al. A new risk prediction model for the assessment of myocardial injury in elderly patients undergoing non-elective surgery. J Cardiovasc Develop Dis. 2024;12:6.
34. Orhan AL, Çınar T, Hayıroğlu Mİ, Çiçek V, Selçuk M, Doğan S, Asal S, Yavuz S, Orhan S, Keser N. Atrial fibrillation as a preoperative risk factor predicts long-term mortality in elderly patients without heart failure and undergoing hip fracture surgery. Rev Assoc Med Bras. 2021;67:1633–8.
35. Troisi F, Guida P, Vitulano N, Argentiero A, Passantino A, Iacoviello M, Grimaldi M. Clinical complexity of an Italian cardiovascular intensive care unit: the role of mortality and severity risk scores. J Cardiovasc Med (Hagerstown). 2024;25:511–8.

36. Guo QY, Peng J, Shan TC, Xu M. Risk factors for mortality in critically ill patients with coagulation abnormalities: a retrospective cohort study. Curr Med Sci. 2024;44:912–22.
37. Li P, Wang R, Liu F, Ma L, Yang H, Qu M, Liu S, Sun M, Liu M, Ma Y, Mi W. High body mass index is associated with elevated risk of perioperative ischemic stroke in patients who underwent noncardiac surgery: a retrospective cohort study. CNS Neurosci Ther. 2024;30: e14838.
38. Rigo-Bonnin R, Gumucio-Sanguino VD, Pérez-Fernández XL, Corral-Ansa L, Fuset-Cabanes M, Pons-Serra M, Hernández-Jiménez E, Ventura-Pedret S, Boza-Hernández E, Gasa M, et al. Individual outcome prediction models for patients with COVID-19 based on their first day of admission to the intensive care unit. Clin Biochem. 2022;100:13–21.

Publisher's Note