scientific reports

Check for updates

OPEN

Development and validation of machine learning-based prediction model for outcome of cardiac arrest in intensive care units

Peifeng Ni[1,2], Sheng Zhang[3], Gensheng Zhang[4], Weidong Zhang[2,5], Hongwei Zhang[2], Ying Zhu[2], Wei Hu[2]✉ & Mengyuan Diao[1,2]✉

Cardiac arrest (CA) poses a significant global health challenge and often results in poor prognosis. We developed an interpretable and applicable machine learning (ML) model for predicting in-hospital mortality of CA patients who survived more than 72 h. A total of 721 patients were extracted from the Medical Information Mart for Intensive Care IV database, divided into the training set (*n* = 576) and the internal validation set (*n* = 145). The external validation set containing 856 cases were collected from four tertiary hospitals in Zhejiang Province. The primary outcome was in-hospital mortality. Eleven ML algorithms were utilized to establish prediction models based on data from 72 h after return of spontaneous circulation (ROSC). The results indicate that the CatBoost model exhibited the best performance at 72 h. Eleven variables were ultimately selected as key features by recursive feature elimination (RFE) to construct a compact model. The final model achieved the highest AUC of 0.86 (0.80, 0.92) in the internal validation and 0.76 (0.73, 0.79) in the external validation. SHAP summary plots and force plots visually explained the predicted outcomes. In conclusion, 72-h CatBoost showed promising performance in predicting in-hospital mortality of CA patients who survived more than 72 h. The model still requires further optimization and improvement.

Keywords Cardiac arrest, Mortality, Machine learning, Categorical boosting, SHapley additive explanations, MIMIC-IV database

Cardiac arrest (CA) is a life-threatening condition, constituting a significant global disease burden. Nearly 380,000 deaths resulting from CA of any cause are reported annually in the United States[1]. Despite advancements in medical treatment, CA still exhibits high rates of death and disability, often linked to post-cardiac arrest brain injury and other forms of fatal organ dysfunction[2,3].

Mortality prediction is crucial for clinicians to identify high-risk factors and intervene promptly, thereby enhancing the outcome of CA patients. Although general severity of illness scores, such as Sequential Organ Failure Assessment (SOFA), Acute Physiological and Chronic Health Evaluation (APACHE) II, and Simplified Acute Physiology Score (SAPS) can be used to predict the mortality of CA patients, they demonstrate only moderate discrimination[4–7]. Despite the development of specific risk scores, obvious limitations exist. The NULL-PLEASE score was an effective predictor of in-hospital mortality[8,9], while the CREST score was only used to predict the mortality of patients with non-ST-segment elevation myocardial infarction[10]. Both scores apply only to out-of-hospital cardiac arrest (OHCA) and involve pre-hospital variables, which are often inaccurately recorded[11]. Additionally, most scores utilize traditional regression analysis for simple variable weighting.

The burgeoning growth of big data and the establishment of medical databases are fostering the integration of machine learning (ML) into clinical practice[12]. ML, learning from extensive data via computers, develops algorithms and constructs prediction models, enhancing computational speed and prediction accuracy[13].

[1]Department of Critical Care Medicine, Zhejiang University School of Medicine, No. 866 Yuhangtang Road, Hangzhou 310000, Zhejiang, China. [2]Department of Critical Care Medicine, Hangzhou First People's Hospital, Westlake University School of Medicine, No. 261 Huansha Road, Hangzhou 310000, Zhejiang, China. [3]Department of Critical Care Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin 2nd Road, Shanghai 200000, China. [4]Department of Critical Care Medicine, Second Affiliated Hospital, Zhejiang University School of Medicine, No. 88 Jiefang Road, Hangzhou 310000, China. [5]Department of Critical Care Medicine, The Fourth Clinical School of Zhejiang Chinese Medicine University, No. 548 Binwen Road, Hangzhou 310000, Zhejiang, China. ✉email: huwei@hospital.westlake.edu.cn; diaomengyuan@hospital.westlake.edu.cn

Previous studies have applied ML to predict the mortality of CA patients. Nanayakkara et al. developed a ML model to predict in-hospital mortality based on a cohort of nearly 40,000 OHCA patients, and they found its performance significantly superior to disease scores[14]. Wong et al. introduced a SARICA score using an interpretable ML framework to predict the survival rate of OHCA[15]. Cheng et al. combined baseline comorbidities and clinical variables to build ML models for survival-to-discharge rate, with eXtremely Gradient Boosting (XGBoost) yielding the best performance[16]. Although ML appears to be a reliable tool for developing prediction models, some studies lack further external validation[14–16]. Furthermore, most studies concentrate on the characteristics within 24 h after admission to assess the severity of CA patients at the earliest stage, while few studies have explored the predictive value of data at 72 h after ROSC. Current guidelines recommend a prognosis assessment at 72 h after rewarming from target temperature management (TTM)[17,18]. Early data may not fully reflect the dynamic changes in the pathophysiological status of patients, and are susceptible to destabilizing factors such as drugs and hypothermia. This underscores the importance of considering the time specificity of clinical variables when constructing prediction models to avoid misjudging adverse outcomes and reducing the survival rate.

In this study, we studied CA patients with survival more than 72 h, evaluated the performance of several ML algorithms, and used variables at 72 h after ROSC to develop a time-specific model in predicting in-hospital mortality for CA. Additionally, we provided visual interpretation to help clinicians fully comprehend the outcomes. Finally, the model performance was demonstrated in external validation.

Methods

Patient cohorts

Data were sourced from the Medical Information Mart for Intensive Care IV database (MIMIC-IV) version 2.2, an open-access critical care database comprising over 70,000 intensive care unit (ICU) stays admitted to the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019[19]. The database encompasses various high-quality data, including demographic information, vital signs, laboratory tests, medications and procedures, diagnostic codes, and illness severity scores, all freely available. As the health data in the MIMIC-IV database is deidentified, the Institutional Review Board at the BIDMC granted a waiver of informed consent.

Patients with CA from any cause were included, defined by ICD-9 codes of 4275 or ICD-10 codes of I46. Exclusion criteria were as follows: (1) Age < 16 years; (2) Length of ICU stay < 72 h. Additionally, a single patient could have multiple ICU stays during hospitalization, with only the first ICU stay being included. The exclusion criteria for length of ICU stay less than 72 h was emphasized, namely the survival time required for the study was greater than 72 h (patients who recovered within 72 h and were transferred to the ward were also excluded due to incomplete data). Patients were randomly divided into the training set (80%) and the internal validation set (20%).

The external validation cohort comprised patients from four tertiary hospitals in Zhejiang Province, namely, Hangzhou First People's Hospital, the Second Affiliated Hospital Zhejiang University School of Medicine, Huzhou Central Hospital, and Jinhua Central Hospital. The inclusion and exclusion criteria mirrored those of the MIMIC cohort. Approval for this study was obtained from the Ethics Committee of Hangzhou First People's Hospital (IIT-20230420-0077-01). We confirmed that all research procedures have been performed in accordance with the Declaration of Helsinki.

The flow chart of the research design is shown in Fig. 1.

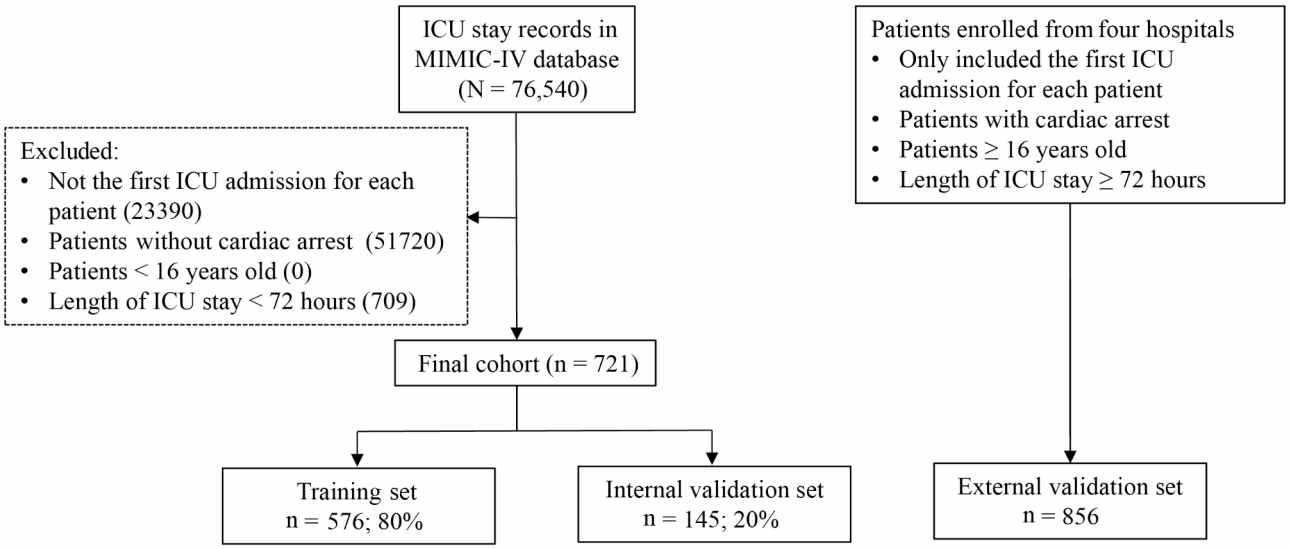

Fig. 1. Flow chart of patient selection.

Data extraction and outcome definition

Based on previous literature[14,16,20,21] and expert opinions, a total of 34 variables were collected (See Table 1 for details). Structured query language with PostgreSQL was employed to extract data from the MIMIC-IV database, encompassing demographic characteristics, vital signs, laboratory results, clinical scores, and treatment measures. ICD-9/10 codes were utilized to identify comorbid conditions. Vital signs and laboratory indicators were recorded 72 h after ROSC. Necessary calculations were performed on relevant data, such as calculating averages (e.g., blood pressure) or sum values (e.g., input) over a interval for variables involving multiple measurements. Comorbidities and treatment measures were transformed into categorical variables.

The primary outcome variable in this study was patient in-hospital mortality, presented as the vital status of CA patients at discharge, indicating whether they survived or died.

Data preprocessing

Addressing missing values is crucial to prevent analytical issues and biased results. The details of missing values were shown in Table S1. In this study, variables with more than 25% of missing values were excluded. Multivariate imputation, a widely utilized method for handling missing values, generates multiple predictions for each missing value based on observed values and estimates the distribution through model building and

Variables	Survival ($n=402$)	Death ($n=319$)	P-value
Male (n, %)	265 (65.9)	191 (59.9)	0.11
Age (years, Median (Q1, Q3))	65 (54, 77)	69 (56, 80)	0.11
Comorbidity (n, %)			
Hypertension	282 (70.2)	208 (65.2)	0.18
Heart failure	164 (40.8)	101 (31.6)	0.01
Cerebral infarction	36 (8.9)	45 (14.1)	0.04
Chronic obstructive pulmonary disease	56 (13.9)	59 (18.5)	0.12
Cirrhosis of liver	11 (2.7)	16 (5.0)	0.16
Chronic kidney disease	101 (25.1)	75 (23.5)	0.68
Malignant cancer	35 (8.7)	40 (12.5)	0.12
Vital signs (Median (Q1, Q3))			
Heart rate (bpm)	83.0 (74.1, 93.8)	89.5 (77.2, 100.5)	<0.001
Respiratory rate (bpm)	19.3 (16.8, 22.2)	19.9 (17.2, 23.1)	0.07
SBP (mmHg)	117.5 (108.5, 130.4)	116.0 (105.1, 129.4)	0.05
DBP (mmHg)	61.1 (54.6, 68.6)	58.7 (52.7, 66.6)	0.01
MBP (mmHg)	78.1 (71.2, 85.6)	75.8 (69.1, 84.7)	0.02
Fluid input (mL)	2297.5 (1362.5, 3595.0)	2700.0 (1577.5, 4340.0)	0.001
Laboratory parameters (Median (Q1, Q3))			
Hemoglobin (g/dL)	10.0 (8.9, 11.3)	9.8 (8.7, 11.1)	0.13
WBC (10^9/L)	10.4 (8.1, 13.7)	13.0 (8.9, 17.7)	<0.001
Platelets (10^9/L)	144.8 (107.0, 195.8)	157.0 (101.2, 214.7)	0.19
Creatinine (mg/dL)	1.2 (0.8, 2.1)	1.5 (0.9, 3.2)	0.001
Glucose (mg/dL)	119.0 (102.0, 148.4)	127.0 (109.3, 162.0)	0.002
Bicarbonate (mmol/L)	24.0 (21.0, 27.0)	22.0 (19.0, 25.0)	<0.001
Sodium (mmol/L)	138.5 (136.0, 141.3)	139.0 (135.5, 142.5)	0.14
Potassium (mmol/L)	4.0 (3.8, 4.4)	4.2 (3.8, 4.7)	<0.001
Chloride (mmol/L)	104.0 (100.0, 108.0)	105.7 (101.0, 110.0)	0.002
Scoring systems (Median (Q1, Q3))			
SOFA	6 (4, 8)	7 (5, 10)	<0.001
Treatment (n, %)			
Percutaneous coronary intervention	24 (6.0)	14 (4.4)	0.44
Extracorporeal membrane oxygenation	8 (2.0)	8 (2.5)	0.83
Continuous renal replacement therapy	43 (10.7)	50 (15.7)	0.06
Intra aortic balloon pump	24 (6.0)	14 (4.4)	0.44
Ventilation	397 (98.8)	317 (99.4)	0.47
Use of vasoactive	292 (72.6)	245 (76.8)	0.24
Use of antiarrhythmic	126 (31.3)	88 (27.6)	0.31
Use of glucocorticoids	28 (7.0)	22 (6.9)	1
Use of sodium bicarbonate	96 (23.9)	115 (36.1)	<0.001

Table 1. Baseline characteristics of patients from the MIMIC-IV database. WBC, white blood cell; SBP, systolic blood pressure; DBP, diastolic blood pressure; MBP, mean blood pressure.

iteration[22]. We employed multivariate imputation via the 'mice' package in R software to conduct the remaining missing values. Predictive mean matching imputation method was chosen to generate five imputed datasets. Each imputed dataset underwent regression analysis, and the results were then combined based on Rubin's rules to obtain a more stable and reliable final estimate. Additionally, we utilized one-hot encoding to transform categorical data, ensuring compatibility with various algorithms.

Development and validation of the model

Before modeling, all of the collected data underwent statistical analysis. Continuous variables were presented as means [standard deviation (SD)] or medians [interquartile range (IQR)]. Student's t test and Wilcoxon Rank Sum test were utilized to compare normally distributed and non-normally distributed data, respectively. Categorical variables were presented as total numbers with percentages and analyzed using Chi-Square tests. $P < 0.05$ was considered to be statistically significant.

Data were randomly divided into the training set (80%) and the internal validation set (20%). Synthesizing Minority Oversampling Technology (SMOTE) was applied to preprocess the training, addressing the issue of unbalanced positive and negative samples and enhancing model robustness. We selected common ML algorithms known for their performance in previous studies[13,23,24], including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Random Forest (RF), AdaBoost, Gradient Boosting Decision Tree (GBDT), eXtremely Gradient Boosting (XGBoost), CatBoost, LightGBM, and Multi-Layer Perceptron (MLP). Based on the variables at 72 h after ROSC, the above algorithms were used to build models in predicting the in-hospital mortality for CA patients. The model demonstrating the best performance was identified as the optimal model. All of the models were developed using Python (version 3.9.7).

To eliminate invalid variables with redundant information, we employed Recursive Feature Elimination (RFE) to identify the key features. Multicollinearity between variables was diagnosed by calculating the Variance Inflation Factor (VIF). The retained features of lower dimensions serve to reduce model complexity, and the resulting model was termed the "compact model". Subsequently, hyperparameter tuning was conducted through random searches, involving 100 trials to enhance model accuracy. The optimized hyperparameters included Learning Rate (learning_rate), L2 Leaf Regularization (l2_leaf_reg), Tree Depth (depth), Bagging Temperature (bagging_temperature), and Minimum Data in Leaf (min_data_in_leaf).

The model performance primarily hinged on the area under the receiver operating characteristic (AUC). Additional evaluation metrics, including sensitivity, specificity, accuracy, F1 score, and Youden's index, were also calculated. We applied the same approach to assess model performance during external validation. Furthermore, calibration plots and decision curve analysis (DCA) were utilized to reflect the actual clinical efficacy of the model.

Given the inherent "black box" nature of ML algorithms, the interpretability of models is often lacking. SHapley Additive exPlanations (SHAP) is a popular method for elucidating complex relationships between features and predictions[25]. SHAP evaluates the importance of each feature based on Shapley values, revealing the contribution of specific features to the given outputs. SHAP summary plots and force plots are employed for global and individual interpretation, respectively. This allows clinicians to intuitively comprehend the underlying process of a particular prediction, aiding in the identification of risk factors and timely intervention.

All of the aforementioned statistical processes were executed using Python (version 3.9.7) in conjunction with R (4.1.2).

Result

A total of 721 admissions involving patients who experienced CA and survived more than 72 h were ultimately identified from the MIMIC-IV database. This cohort was subsequently randomly split into the training set (80%, $n = 576$) and the internal validation set (20%, $n = 145$). Additionally, the external validation cohort comprised 856 eligible patients. Table 1 and Table S2 displayed all of the included variables at 72 h after ROSC.

Baseline characteristics

As depicted in Table 1, we conducted a comparison of baseline characteristics between the survival and death groups. The in-hospital mortality of CA patients in the MIMIC cohort was 44.24% (319 deaths and 402 survivors). Some variables were considered to be significant between the groups, including vital signs such as heart rate, diastolic blood pressure (DBP), and mean blood pressure (MBP); laboratory parameters such as white blood cells (WBC), potassium, chloride, bicarbonate, glucose, and creatinine; fluid input; SOFA score; use of sodium bicarbonate; and comorbidities such as cerebral infarction and heart failure.

Model development and validation

We constructed prediction models based on 11 ML algorithms, employing all of the available variables at 72 h after ROSC. Figure 2 illustrates the comparison of the performance of all algorithms at 72 h. CatBoost demonstrated superior discriminative power compared to other algorithms, with an AUC (95% CI) of 0.84 (0.78, 0.91). As a result, CatBoost was chosen as the algorithm for modeling.

To streamline the model, 11 variables were ultimately selected as key features through RFE, including age, heart rate, DBP, MBP, WBC, platelets, creatinine, sodium, chloride, bicarbonate, and heart failure. After 100 trials, the optimal hyperparameters of the CatBoost model were determined (Table S3 and Figure S1). Figure 3A demonstrates that the compact CatBoost model, incorporating the aforementioned 11 variables at the 72-h mark, generated the best performance among various ML algorithms, with an AUC of 0.86 (0.80, 0.92), sensitivity of 0.73 (0.62, 0.82) and specificity of 0.88 (0.77, 0.94). In addition to ROC curves, Table S4 provides a comprehensive overview of the evaluation indicators for each model. CatBoost achieved the highest Youden's index, despite not leading in every other metric. The compact model outperformed the full model (AUC: 0.86 vs. 0.84), even with

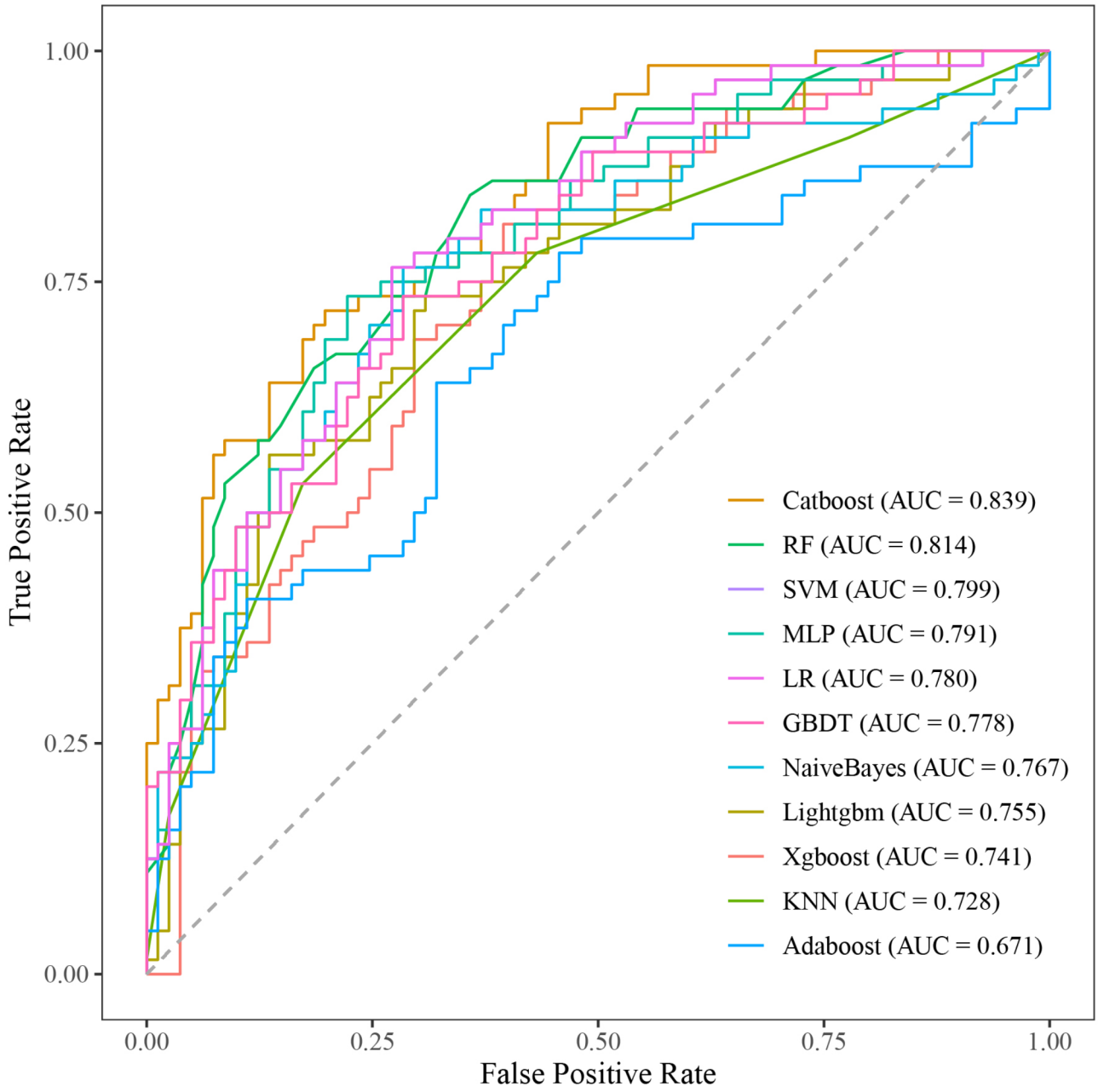

Fig. 2. Comparison of the model performance based on different algorithms. RF, Random Forest; SVM, Support Vector Machine; MLP, Multi-Layer Perceptron; LR, Logistic Regression; GBDT, Gradient Boosting Decision Tree; XGBoost, eXtremely Gradient Boosting; KNN, K-Nearest Neighbor; AUC, areas under receiver operating characteristic curves.

the removal of some variables (Fig. 3B). The comparison between the CatBoost model and the SOFA score was depicted in Fig. 3C, clearly highlighting the superiority of the 72-h CatBoost model (AUC: 0.86 vs. 0.65). Delong test demonstrated that the differences in predictive performance between the compact model and the full model as well as the SOFA score were statistically significant, with *P* values of 0.03 and 0.005, respectively.

The calibration plot presents the agreement between estimated and observed probabilities (Fig. 4A), and the analysis details of each bin are shown in Table S5. The DCA demonstrates that the 72-h CatBoost model can guide decision-making and achieve clinical benefit across a wide range of threshold probabilities (Fig. 4B), which further proves the feasibility of application to clinical practice. Therefore, we believe that the 72-h CatBoost model had a preferable discrimination ability.

Interpretability analysis

SHAP methods were employed to provide a visual interpretation of the model outputs. Initially, a global interpretation of the compact model was carried out by utilizing the SHAP summary plot. Feature weights were

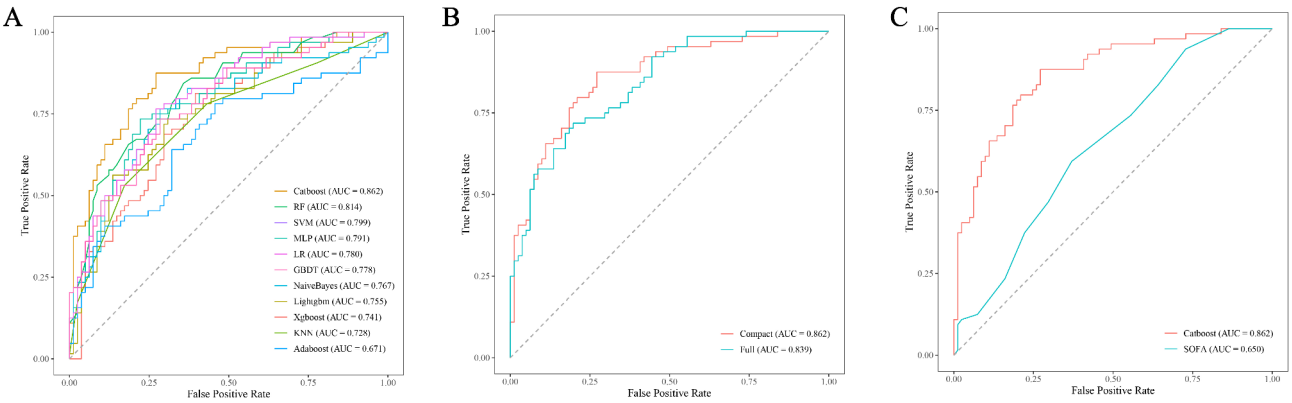

Fig. 3. Comparison of the model performance in internal validation. (**A**) ROC of all of the models. (**B**) ROC full model vs. compact model. (**C**) ROC compact model vs. SOFA.

represented as SHAP values revealing the overall impact of each feature on the prediction of mortality, ordered by the sum of absolute SHAP values across all of the participates. As illustrated in Fig. 5, this plot also facilitated feature importance analysis. The top five effective features in the compact model for predicting the mortality of CA were WBC, heart rate, bicarbonate, MBP, and age. Each dot in the plot represents one patient, with the colors ranging from blue to red indicating the attribution value—blue corresponding to a lower feature value and red to a higher feature value. Moreover, the greater the distance of a colored dot from the centerline (baseline SHAP value of zero), the more pronounced its impact on the result.

SHAP force plots were illustrated to provide detailed and individual interpretations of model predictions. They present the mortality risk of specific patients, showcasing the contribution level of each characteristic. These plots are particularly valuable to help clinicians comprehend specific outputs and identify critical risk factors, thereby assisting in decision-making. Figure 6 illustrates two examples. In a single case, each variable is represented by an arrow, with red indicating a positive effect toward death, and blue for a negative effect. The size of the arrow corresponds to the importance of the variable in predicting mortality for that specific case, with larger arrows signifying a more substantial impact. The combined effects of all variables contribute to the predicted output value of the model. Figure 6 shows a case of death and a case of survival. In Fig. 6A, based on the model prediction, the final output predicted value for death was 1.54. MBP at 72 h acted as a protective factor, while bicarbonate, WBC, age, creatinine, chlorine, and heart rate greatly increased the risk of death in this case. Among these factors, bicarbonate at 72 h emerged as the most critical and threatening contributor.

External validation

Data from a total of 856 eligible CA patients from four tertiary hospitals in Zhejiang Province were collected to validate the model. The in-hospital mortality rate was 50.7% (434 deaths and 422 survivors). A comparison between the external validation cohort and the MIMIC cohort is presented in Table S6. Notably, CA patients in the external validation cohort exhibited a higher rate of in-hospital mortality compared to the MIMIC cohort. In addition, the external validation cohort had a younger age, higher heart rate, blood pressure, bicarbonate, sodium, and chlorine, and lower WBC, platelets, and creatinine.

The validation results, as displayed in Fig. 7, revealed that the 72-hour CatBoost model calculated an AUC (95% CI) of 0.76 (0.73, 0.79), significantly outperforming other models. Additional metrics are presented in Table S7. Calibration curves and decision curves for external validation are shown in Figure S2.

Discussion

In this study, we developed ML models for predicting the in-hospital mortality of CA patients who survived more than 72 h. We screened 11 key variables, and the 72-hour CatBoost model achieved the highest predictive ability. The SHAP method facilitated both global and local interpretations of the model, further enhancing its transparency. Additionally, the model exhibited robust discriminative capabilities in external validation.

In recent years, ML has played an unparalleled role in early warning systems[26,27], decision support, and subphenotypes identification of CA[28,29], promisingly contributing to the reduction of disease burden. Clinicians stand to benefit from a deeper understanding of the prognosis of CA patients, enabling timely treatment adjustments to prevent unfavorable outcomes. We introduced the CatBoost algorithm and discovered its efficacy in evaluating in-hospital mortality for CA patients. CatBoost, an innovative GBDT algorithm, is designed to automatically handle ML studies involving categorical variables[30]. Researchers have emphasized the crucial role of hyperparameter optimization in maximizing the efficiency of CatBoost[30]. CatBoost has shown satisfactory predictive power in previous studies. Li et al. predicted the in-hospital mortality for mechanically ventilated patients with congestive heart failure[24], and Chen et al. predicted the 90-day prognosis of patients with transient ischemic attack[31]. To date, CatBoost has not been used for modeling in any CA studies. Notably, our study is the first to demonstrate the practicality of CatBoost in predicting in-hospital mortality of CA patients.

Traditionally, studies have primarily depended on clinical variables within the first 24 h to predict the mortality of CA patients, and rarely explored the prognostic value of variables at 72 h after ROSC. Both

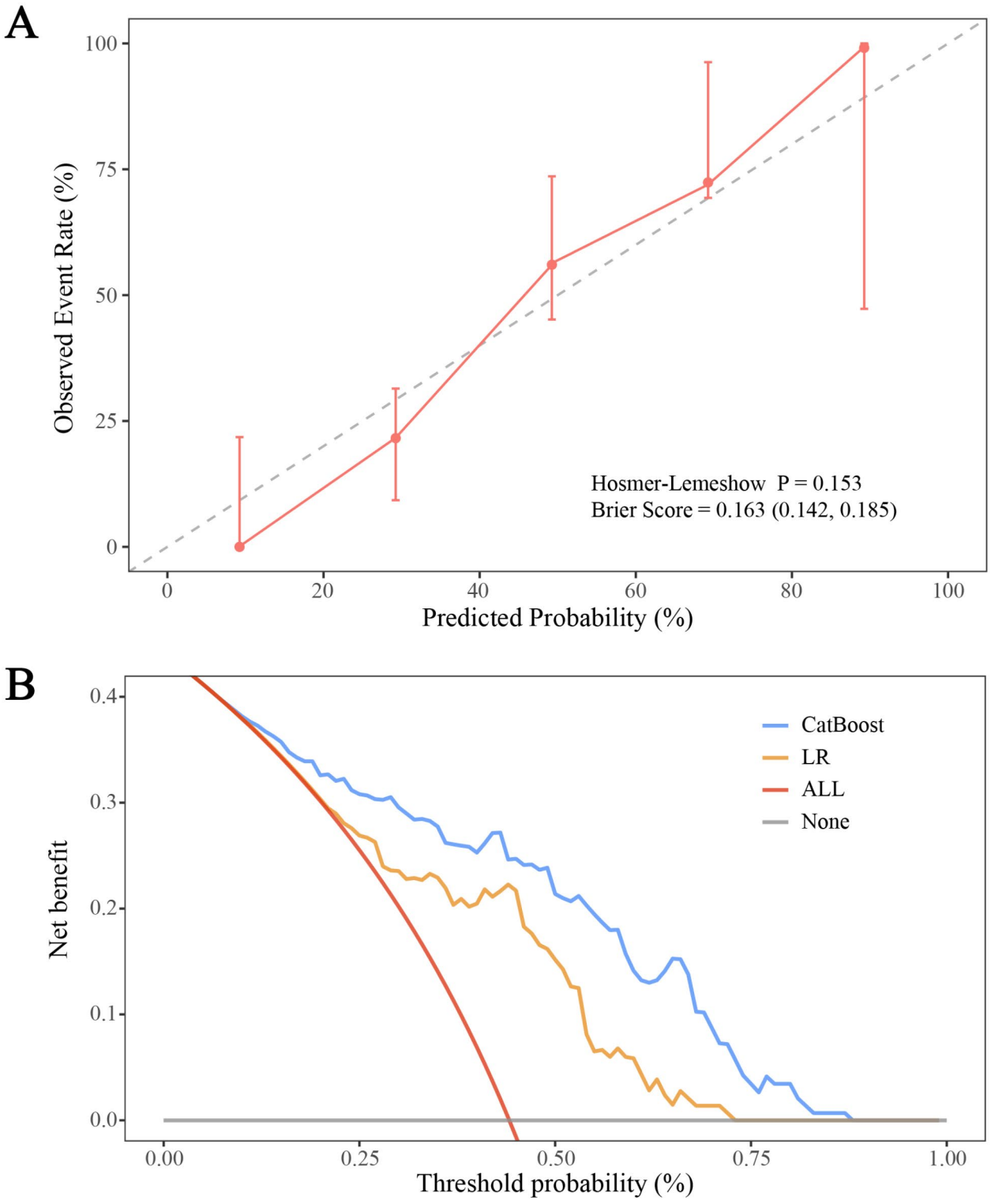

Fig. 4. Clinical practicability of the model. (**A**) Calibration plot of the CatBoost model and LR. (**B**) Decision curve analysis of the CatBoost model and LR.

European and American guidelines recommend multimodal prognostic assessment at 72 h after TTM[17,18]. Besides, an important concern is that for patients with extended survival times, 24-hour or 48-hour data may not fully capture the dynamic changes in their pathophysiological conditions, leading to a risk of false positive rates and an overestimation of predicted mortality. Our study focused on CA patients who survived more than 72 h after ICU admission, using 72-hour data to construct a model in predicting in-hospital mortality for these patients. Since we excluded patients who died within 72 h, this cohort showed higher survival rates.

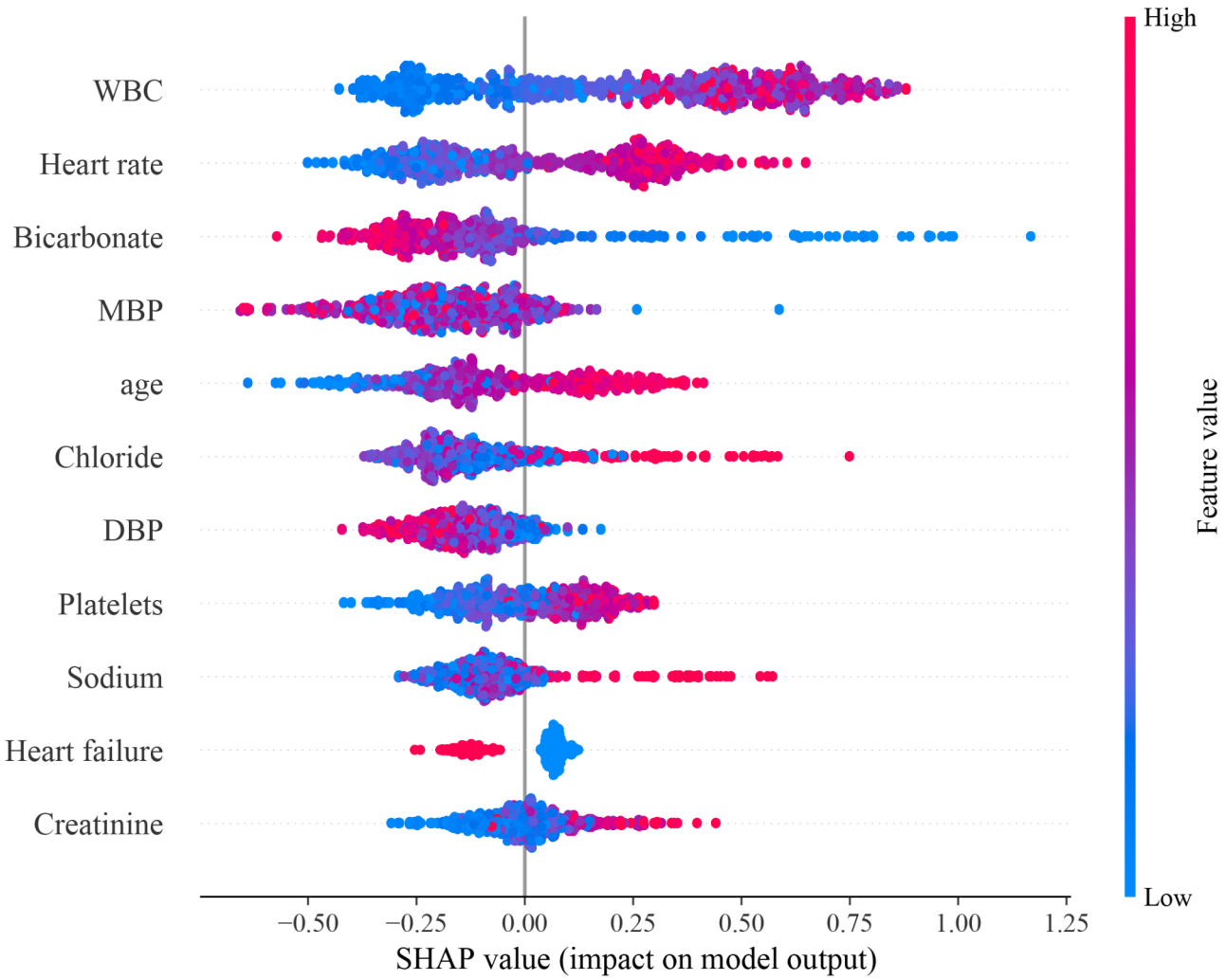

Fig. 5. SHAP summary plot of the model. The plot was based on 72-hour variables. The colors of dots ranging from blue to red represent the attribution value, where blue corresponds to a lower feature value and red to a higher value. According to the overall contribution value, the importance of the variables was ranked as WBC, heart rate, bicarbonate, MBP, age, chlorine, DBP, platelets, sodium, heart failure, and creatinine.

The temporal dynamics of vital signs and laboratory results allow the most recent data more reflective of the actual physiological state of CA patients, thereby enhancing its predictive significance. For example, researchers using APACHE II to predict the in-hospital mortality and neurological outcome of CA patients found that the discrimination ability of the score increased over time, reaching its peak at 72 h[5]. Andersson et al. undertook a similar approach by developing artificial neural network models based on clinical variables and biomarkers. Their study corroborated our observation, revealing that the model achieved the best predictive performance on the third day (AUC: 0.941)[32]. This underscores the importance of considering time specificity in constructing prediction models to avoid misjudging adverse outcomes and making premature decisions regarding withdrawal of life-sustaining therapy.

We identified key predictors of CA mortality and found support in previous reports. Age, commonly associated with mortality[33], particularly in the elderly who face an increased risk of comorbidities such as heart failure, emerged as a significant factor[34]. Abnormal hemodynamics, reflected in lower blood pressure and higher heart rate, concurred with previous research emphasizing hypotension and tachycardia as independent indicators of mortality[35–38]. Leukocytosis is the consequence of inflammation caused by tissue ischemia-reperfusion injury post-CA with secondary infection[39]. WBC emerged as the strongest independent predictor in our study, emphasizing its role in mortality prediction. The occurrence of hypoperfusion in CA patients initiates a cascade of physiological responses leading to metabolic acidosis and kidney insult. This results in disturbances such as water-electrolyte imbalance and renal insufficiency[40], which are reflected in deviations from normal levels of sodium, chlorine, bicarbonate, and creatinine. Our final prediction model exclusively incorporated these clinically common indicators, enhancing the model's practicality. The removal of insignificant variables did not compromise the model's predictive performance.

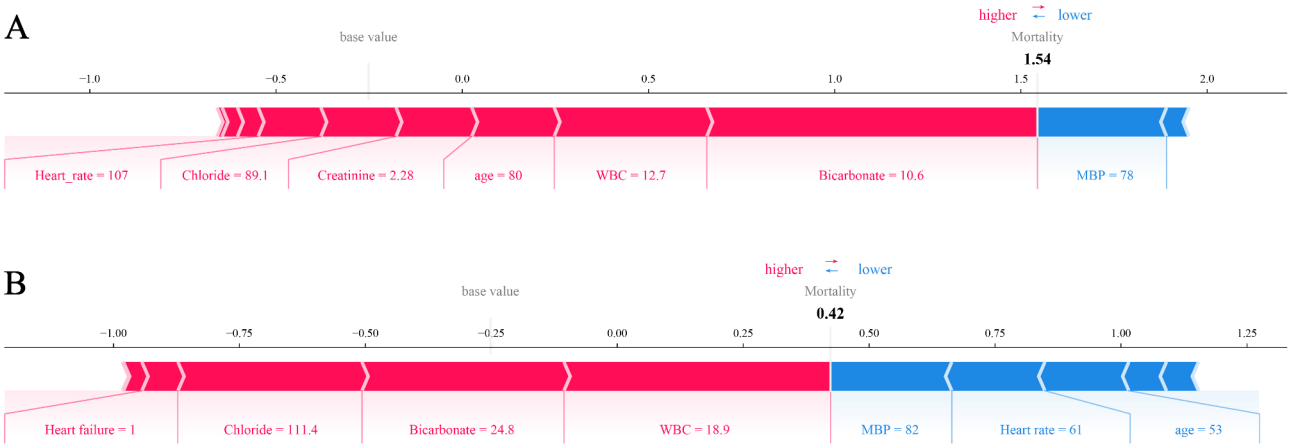

Fig. 6. SHAP force plots of the model. The plots were based on 72-h variables. Each variable is represented by an arrow, with red indicating a positive effect on death and blue a negative effect. The size of the arrow indicates the importance of the variable in predicting mortality, where a larger arrow indicates a greater impact. The combined effects of all of the variables contribute to the predicted output value of the model. (**A**) The patient died at discharge with an output predictive value of 1.54. (**B**) The patient survived at discharge with an output predictive value of 0.42.

Recognizing the inherent black-box nature of ML algorithms, we introduced SHAP to visualize the intrinsic connection between input and response, making complex risk predictions more interpretable[25]. SHAP summary plots provided insights into the overall distribution of input variables' contributions to predicted outcomes, revealing the order of importance of each feature. Based on SHAP values, the top five predictors were WBC, heart rate, bicarbonate, MBP, and age. These indicators reflect the immunity, circulation and metabolism function of patients, with their deterioration greatly contributing to adverse outcomes. SHAP force plots clearly illustrate the internal formation mechanism of individual prediction outcomes. Calculated mortality risk and visualized variable effects facilitate interaction with clinicians. Thus, clinicians can gain a comprehensive understanding of the personalized risk profiles of CA patients. By accurately detecting risk factors and abnormal indicators through SHAP analysis, clinicians can optimize treatment strategies to suit individual needs.

Most importantly, our study provided external validation of the self-developed model for predicting in-hospital mortality of CA. Although several predictive models in previous studies have demonstrated excellent performance, only a few have undergone external validations[14–16]. Ours is the first study in which multi-center external validation was performed using data from Chinese hospitals. Although our model demonstrated decent discriminability in external validation, its overall performance was inferior to that in internal validation. We analyzed the reasons for this discrepancy. First and foremost, there are notable differences in the distribution of data across different countries, as shown in Table S6, thereby directly affecting the model's generalizability. Secondly, some variables have a high proportion of missing values in the external validation set, such as bicarbonate (Table S1); we employed multiple imputation to address these missing values, but the authenticity of the data may have been compromised. Since any missing value could impede the implementation of the model, the prerequisite is to ensure the completeness of input data in clinical settings. Additionally, retrospective data collection may introduce information bias. Furthermore, the model may have become overfitted to the training set. Therefore, we need to strengthen quality control during retrospective data collection to enhance data integrity and accuracy, and conduct further validation in more diverse cohorts to improve the model's generalizability and robustness.

However, it is crucial to acknowledge the limitations of our current study. First, the absence of pre-hospital data and some specific biomarkers represent a gap, such as cause of CA, no-flow time, initial rhythm, neuron-specific enolase (NSE), etc. Given the unavailability of these data from the MIMIC database and the retrospective nature of our study, acquiring comprehensive data in these domains is challenging and unrealistic. Second, our study built the prediction model based on variables at 72 h. This is not a universal model; it is not intended for early (within the first 24 h) mortality prediction but rather for providing prognostic assessment for CA patients who survived more than 72 h. Thirdly, our cohorts remain small-scale datasets with fewer than 1000 cases, so further validation of the model's generalizability within larger cohorts is warranted. Finally, being a retrospective study, future prospective studies are needed to validate the model in real-time clinical settings.

In conclusion, ML is an effective tool for predicting in-hospital mortality of CA patients. The 72-hour CatBoost model proved to be accurate and valid in external validation. The integration of ML into the prognostic assessment of CA proves crucial for clinicians in optimizing treatment decisions, formulating precise management strategies, and maximizing the survival rate of CA survivors.

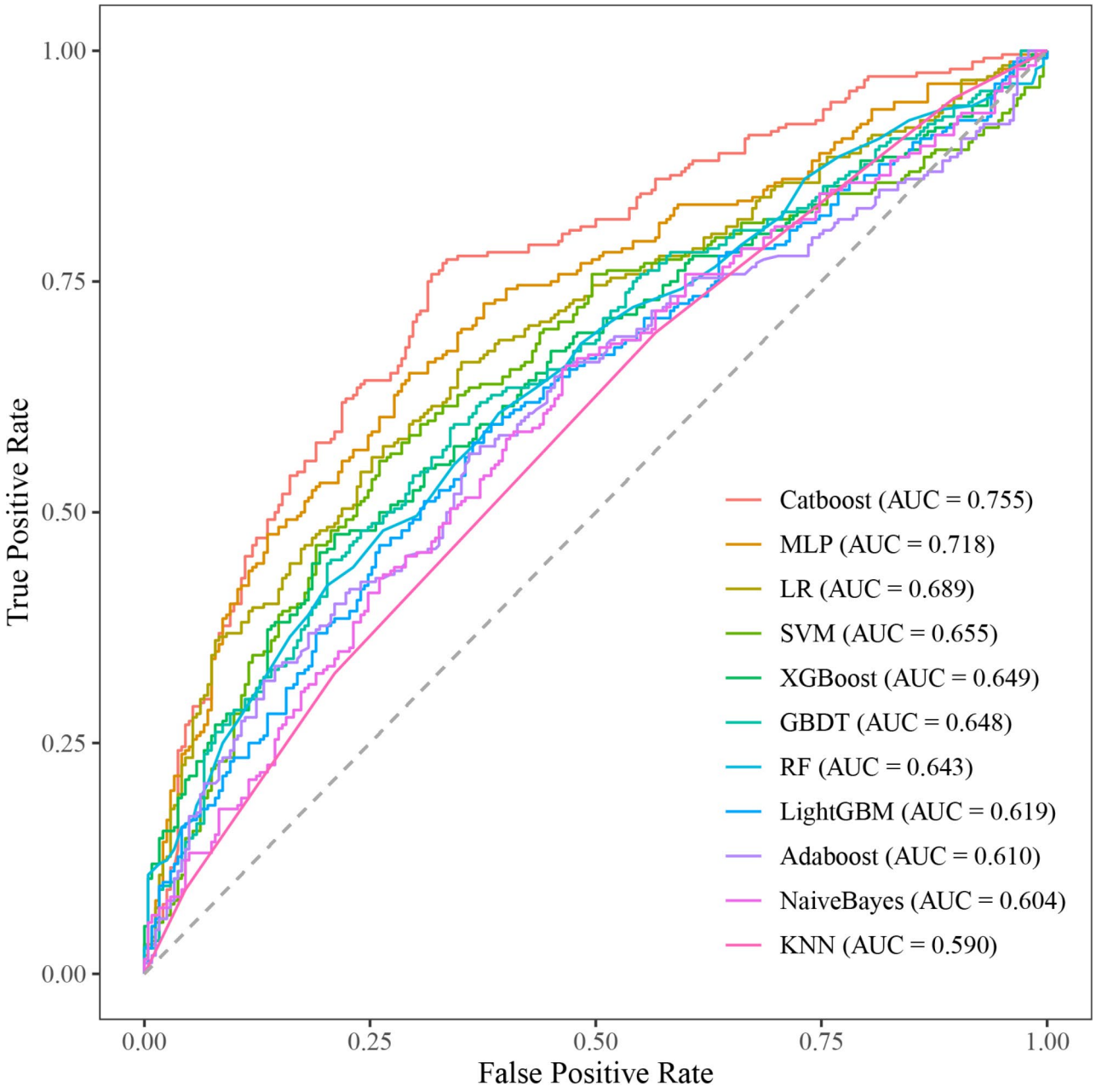

Fig. 7. Model performance in the external validation.

Data availability

The datasets supporting the conclusions of this article are available in MIMIC-IV (https://doi.org/10.13026/7vcr-e114) and are available from the corresponding author on reasonable request.

Received: 8 September 2024; Accepted: 5 March 2025
Published online: 13 March 2025

References

1. Virani, S. S. et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation* **141**, E139-E596 (2020). https://doi.org/10.1161/cir.0000000000000757
2. Sandroni, C., Cronberg, T. & Sekhon, M. Brain injury after cardiac arrest: Pathophysiology, treatment, and prognosis. *Intensive Care Med.* **47**, 1393–1414. https://doi.org/10.1007/s00134-021-06548-2 (2021).
3. Penketh, J. & Nolan, J. P. Post-cardiac arrest syndrome. *J. Neurosurg. Anesthesiol.* **35**, 260–264. https://doi.org/10.1097/ana.0000000000000921 (2023).
4. Matsuda, J. et al. The sequential organ failure assessment (SOFA) score predicts mortality and neurological outcome in patients with post-cardiac arrest syndrome. *J. Cardiol.* **76**, 295–302. https://doi.org/10.1016/j.jjcc.2020.03.007 (2020).

5. Donnino, M. W. et al. APACHE II scoring to predict outcome in post-cardiac arrest. *Resuscitation* **84**, 651–656. https://doi.org/10.1016/j.resuscitation.2012.10.024 (2013).
6. Salciccioli, J. D. et al. Performance of SAPS II and SAPS III scores in post-cardiac arrest. *Minerva Anestesiol.* **78**, 1341–1347 (2012).
7. Choi, J. Y. et al. Performance on the APACHE II, SAPS II, SOFA and the OHCA score of post-cardiac arrest patients treated with therapeutic hypothermia. *PLoS ONE.* **13**, 12. https://doi.org/10.1371/journal.pone.0196197 (2018).
8. Gue, Y. X. et al. Usefulness of the NULL-PLEASE score to predict survival in Out-of-Hospital cardiac arrest. *Am. J. Med.* **133**, 1328–1335. https://doi.org/10.1016/j.amjmed.2020.03.046 (2020).
9. Potpara, T. S. et al. External validation of the simple NULL-PLEASE clinical score in predicting outcome of Out-of-Hospital cardiac arrest. *Am. J. Med.* **130**, 9. https://doi.org/10.1016/j.amjmed.2017.05.035 (2017).
10. Bascom, K. E. et al. Derivation and validation of the CREST model for very early prediction of circulatory etiology death in patients without ST-Segment-Elevation myocardial infarction after cardiac arrest. *Circulation* **137**, 273–282. https://doi.org/10.1161/circulationaha.116.024332 (2018).
11. Naik, R., Mandal, I. & Gorog, D. A. Scoring systems to predict survival or neurological recovery after Out-of-hospital cardiac arrest. *Eur. Cardiol. Rev.* **17**, 1–6. https://doi.org/10.15420/ecr.2022.05 (2022).
12. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA-J. Am. Med. Assoc.* **319**, 1317–1318. https://doi.org/10.1001/jama.2017.18391 (2018).
13. Nwanosike, E. M., Conway, B. R., Merchant, H. A. & Hasan, S. S. Potential applications and performance of machine learning techniques and algorithms in clinical practice: A systematic review. *Int. J. Med. Inf.* **159**, 11. https://doi.org/10.1016/j.ijmedinf.2021.104679 (2022).
14. Nanayakkara, S. et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Med.* **15**, 16. https://doi.org/10.1371/journal.pmed.1002709 (2018).
15. Wong, X. Y. et al. Clinical paper development and validation of the SARICA score to predict survival after return of spontaneous circulation in out of hospital cardiac arrest using an interpretable machine learning framework. *Resuscitation* **170**, 126–133. https://doi.org/10.1016/j.resuscitation.2021.11.029 (2022).
16. Cheng, C. Y., Chiu, I. M., Zeng, W. H., Tsai, C. M. & Lin, C. H. R. Machine learning models for survival and neurological outcome prediction of Out-of-Hospital cardiac arrest patients. *Biomed. Res. Int.* **2021** (8). https://doi.org/10.1155/2021/9590131 (2021).
17. Nolan, J. P. et al. European resuscitation Council and European society of intensive care medicine guidelines 2021: Post-resuscitation care. *Intensive Care Med.* **47**, 369–421. https://doi.org/10.1007/s00134-021-06368-4 (2021).
18. Panchal, A. R. et al. Part 3: Adult basic and advanced life support 2020 American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* **142**, S366–S468. https://doi.org/10.1161/cir.0000000000000916 (2020).
19. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data.* **10**, 9. https://doi.org/10.1038/s41597-022-01899-x (2023).
20. Park, J. H. et al. Prediction of good neurological recovery after out-of-hospital cardiac arrest: A machine learning analysis. *Resuscitation* **142**, 127–135. https://doi.org/10.1016/j.resuscitation.2019.07.020 (2019).
21. Sun, Y. W., He, Z. Y., Ren, J. & Wu, Y. F. Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: A retrospective analysis of MIMIC -IV database based on machine learning. *BMC Anesthesiol.* **23**, 17. https://doi.org/10.1186/s12871-023-02138-5 (2023).
22. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr Res.* **20**, 40–49. https://doi.org/10.1002/mpr.329 (2011).
23. Kwon, J. M. et al. Deep-learning-based out-of-hospital cardiac arrest prognostic system to predict clinical outcomes. *Resuscitation* **139**, 84–91. https://doi.org/10.1016/j.resuscitation.2019.04.007 (2019).
24. Li, L. et al. Prediction of hospital mortality in mechanically ventilated patients with congestive heart failure using machine learning approaches. *Int. J. Cardiol.* **358**, 59–64. https://doi.org/10.1016/j.ijcard.2022.04.063 (2022).
25. Lundberg, S. M. & Lee, S. I. In *31st Annual Conference on Neural Information Processing Systems (NIPS).* (Neural Information Processing Systems (Nips) (2017).
26. Wu, T. T., Lin, X. Q., Mu, Y., Li, H. & Guo, Y. S. Machine learning for early prediction of in-hospital cardiac arrest in patients with acute coronary syndromes. *Clin. Cardiol.* **44**, 349–356. https://doi.org/10.1002/clc.23541 (2021).
27. Javan, S. L., Sepehri, M. M., Javan, M. L. & Khatibi, T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. *Comput. Meth Prog. Biomed.* **178**, 47–58. https://doi.org/10.1016/j.cmpb.2019.06.010 (2019).
28. Okada, Y. et al. Clustering out-of-hospital cardiac arrest patients with non-shockable rhythm by machine learning latent class analysis. *Acute Med. Surg.* **9**, 11. https://doi.org/10.1002/ams2.760 (2022).
29. Harford, S. et al. A machine learning approach for modeling decisions in the out of hospital cardiac arrest care workflow. *BMC Med. Inf. Decis. Mak.* **22**, 9. https://doi.org/10.1186/s12911-021-01730-4 (2022).
30. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: An interdisciplinary review. *J. Big Data.* **7**, 45. https://doi.org/10.1186/s40537-020-00369-8 (2020).
31. Chen, S. D. et al. Machine learning is an effective method to predict the 90-day prognosis of patients with transient ischemic attack and minor stroke. *BMC Med. Res. Methodol.* **22**, 11. https://doi.org/10.1186/s12874-022-01672-z (2022).
32. Andersson, P. et al. Predicting neurological outcome after out-of-hospital cardiac arrest with cumulative information; Development and internal validation of an artificial neural network algorithm. *Crit. Care.* **25**, 12. https://doi.org/10.1186/s13054-021-03505-9 (2021).
33. Andrew, E., Mercier, E., Nehme, Z., Bernard, S. & Smith, K. Long-term functional recovery and health-related quality of life of elderly out-of-hospital cardiac arrest survivors. *Resuscitation* **126**, 118–124. https://doi.org/10.1016/j.resuscitation.2018.03.017 (2018).
34. Hirlekar, G. et al. Comorbidity and survival in out-of-hospital cardiac arrest. *Resuscitation* **133**, 118–123. https://doi.org/10.1016/j.resuscitation.2018.10.006 (2018).
35. Bhate, T. D., McDonald, B., Sekhon, M. S. & Griesdale, D. E. G. Association between blood pressure and outcomes in patients after cardiac arrest: A systematic review. *Resuscitation* **97**, 1–6. https://doi.org/10.1016/j.resuscitation.2015.08.023 (2015).
36. Chi, C. Y. et al. Post-resuscitation diastolic blood pressure is a prognostic factor for outcomes of cardiac arrest patients: A multicenter retrospective registry-based analysis. *J. Intensive Care.* **10**, 39. https://doi.org/10.1186/s40560-022-00631-6 (2022).
37. Li, Z. M., Zhou, D. W., Zhang, S. L., Wu, L. & Shi, G. Z. Association between mean arterial pressure and survival in patients after cardiac arrest with vasopressor support: A retrospective study. *Eur. J. Emerg. Med.* **28**, 277–284. https://doi.org/10.1097/mej.0000000000000787 (2021).
38. Matsumoto, S. et al. Heart rate after resuscitation from Out-of-Hospital cardiac arrest due to acute coronary syndrome is an independent predictor of clinical outcome. *Circ. J.* **84**, 569–576. https://doi.org/10.1253/circj.CJ-19-0836 (2020).
39. Langeland, H. et al. The inflammatory response is related to circulatory failure after out-of-hospital cardiac arrest: A prospective cohort study. *Resuscitation* **170**, 115–125. https://doi.org/10.1016/j.resuscitation.2021.11.026 (2022).
40. Domanovits, H. et al. Acute renal failure after successful cardiopulmonary resuscitation. *Intensive Care Med.* **27**, 1194–1199. https://doi.org/10.1007/s001340101002 (2001).

Acknowledgements
We would like to sincerely thank MIMIC-IV for open access to their database. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Author contributions
P.N.: Conceptualization, Methodology, Formal analysis, Investigation, Writing-Original Draft. S.Z.: Conceptualization, Methodology, Writing-Review & Editing. G.Z.: Conceptualization, Writing-Review & Editing, Supervision. W.Z.: Formal analysis, Data Curation, Writing-Review & Editing. H.Z.: Formal analysis, Data Curation, Writing-Review & Editing. Y.Z.: Writing-Review & Editing, Supervision. W.H.: Writing-Review & Editing, Supervision. M.D.: Writing-Review & Editing, Supervision.

Funding
This work was supported by Key Program Cosponsored by Zhejiang Province and National Health Commission of China (Grant. WKJ-ZJ-2315), Construction Fund of Medical Key Disciplines of Hangzhou (Grant. OO20200485), and Science and Technology Development Project of Hangzhou (Grant. 202204A10).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93182-3.

Correspondence and requests for materials should be addressed to W.H. or M.D.