DOI Link: https://doi.org/10.61586/faAOo
Vol.50, Issue.7, Part.1, July 2025, PP.138-163

Mortality Prediction and Risk Factor Identification in Heart Failure Patients Using 17 SHAP-Interpreted Machine Learning Algorithms

Dan Xu[1], Hongli Liu[2], Jing Yang[3*]

1 Department of Ultrasound, The General Hospital of Western Theater Command, Chengdu 610083, China.

2 Outpatient Department (OPD), The General Hospital of Western Theater Command，Chengdu 610083, China

3 Demonstration Center for Experimental Teaching in Biomedicine, Chengdu Medical College, Chengdu, 610500, China.

* Corresponding author.

E-**mail addresses:** Xd_819@outlook.com (Dan Xu), LHL_808@outlook.com (Hongli Liu), jingyang@cmc.edu.cn (Jing Yang)

Received May 2025 Accepted June 2025 Published July 2025

Abstract:

Background: Heart failure (HF) is a leading cause of morbidity and mortality in intensive care units (ICUs). While echocardiography is routinely used for prognosis, critical predictors distinguishing survivors from non-survivors remain poorly defined across diverse clinical settings. Machine learning (ML) offers potential to improve risk stratification by identifying key mortality determinants.

Methods: In this retrospective study, we analyzed data from two large ICU databases: MIMIC-IV and eICU. Adult patients (>18 years) with HF who underwent echocardiographic evaluation during their ICU stay were included. After feature selection using the Boruta algorithm, we developed and compared 17 ML models to predict in-hospital mortality, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR) and others. Models were internally validated on a test subset of MIMIC-IV data and externally validated on the eICU cohort. Model performance was assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and decision curve analysis (DCA). SHapley Additive exPlanations (SHAP) was applied to identify and rank key mortality predictors.

Results: A total of 3,485 patients were included, with 2,738 from the MIMIC-IV database and 747 from the eICU database. Thirty-two potential predictors were identified through screening using the Boruta algorithm. Receiver operating characteristic (ROC) curves analysis showed that the RF model consistently demonstrated superior performance across all validation sets and achieved the highest AUC (0.940) in the combined MIMIC-IV and eICU datasets. SHAP algorithm analysis revealed that the clinical features with the greatest impact on model outputs were SAPS II score, OASIS score, SOFA score, age, white blood cell count within the first 24 hours, and first 24-hour blood urea nitrogen. Clinical interpretability analysis indicated that the positive contributions (SHAP values) of these features were 0.115, 0.090, 0.084, 0.046, 0.035, and 0.015, respectively, demonstrating the model's good granularity and transparency in individualized prediction.

Conclusions: This study compared 17 prediction models, among which the RF model demonstrated the best predictive performance for mortality in HF patients within the ICU, and may assist clinicians in identifying and managing high-risk HF patients.

Keywords: Heart Failure, ICU, Mortality, Echocardiography, Machine Learning, Prediction Model, SHAP

Introduction

Heart failure (HF), also known as congestive heart failure, represents the terminal progression of cardiovascular diseases, with primary clinical manifestations including dyspnea, fluid retention, and fatigue[1]. Designated as an emerging epidemic in 1997, HF imposes a $108 billion annual economic burden worldwide due to its high morbidity, hospitalization rates, and mortality, and has become a critical public health concern[2, 3]. Studies indicate that global HF cases reached 56.5 million in 2021, predominantly affecting elderly populations, with patients over 75 years old accounting for 20% [3-5]. Notably, HF patients requiring hospitalization often need intensive care unit (ICU) admission for advanced monitoring, yet ICU-admitted HF patients maintain elevated mortality rates[6]. Reported in-hospital mortality reaches 10.6% for ICU-managed HF patients versus 4.0% for general HF patients [7], which underscores the urgent need for early mortality prediction in ICU HF patients to guide clinical decision-making.

Given HF's complex etiology, severe progression, and poor prognosis, recent research prioritizes risk prediction and assessment[8]. As a portable, non-invasive imaging technique without radiation exposure, echocardiography (Echo) has emerged as the gold standard for HF diagnosis and prognosis assessment[9]. However, limitations persist in echo's prognostic utility, including operator dependency, indirect assessment of diastolic function, and diagnostic constraints in specific etiologies (e.g., overemphasis on left ventricular parameters while neglecting right ventricular indices)[10-12]. Integrating demographic characteristics, comorbidities, and multimodal imaging parameters (e.g., continuous maximum scores from advanced imaging techniques), and leveraging these integrated factors through the development of robust identification tools, could provide a more comprehensive approach to address echocardiography's limitations while significantly enhancing the accuracy of heart failure diagnosis and prognosis prediction[11, 13-15]. Therefore, developing precise diagnostic and prognostic tools remains imperative to reduce HF mortality and improve clinical risk stratification.

Recent advances demonstrate artificial intelligence (AI)'s potential in enhancing echocardiographic diagnosis, treatment optimization, and risk

prediction[16-18]. Machine learning (ML), a key subset of AI, processes high-volume multimodal data, outperforming traditional statistical methods in prognostic accuracy and predictive factor identification[19-22]. Emerging studies apply ML to Echo-based mortality prediction in cardiovascular patients[23]. Studies have shown that echo-derived ML models significantly improve sensitivity in 1-year all-cause mortality prediction, demonstrating robust performance in internal/external validations (area under the receiver operating characteristic, AUC: 75%-92%) for automated clinical risk stratification[24, 25]. Nevertheless, despite promising accuracy, few existing ML models undergo independent cohort validation, with limited reports on ICU HF mortality prediction models.

To address the critical challenge of predicting HF mortality in ICU, this study focused on high-risk adult ICU patients with HF diagnosed by echo and comparatively analyzed 17 ML models to develop a mortality prediction model integrating echo assessments, clinical characteristics, and clinical interventions. Leveraging two large independent databases (MIMIC-IV and eICU) to obtain representative data, we employed the Boruta algorithm for objective feature selection, compared 17 ML models to mitigate selection bias, and ensured model robustness through internal and external validation. Additionally, SHAP (SHapley Additive exPlanations) analysis was applied to quantify the contribution of predictive variables, systematically advancing early identification and prognostic improvement for high-risk HF patients.

Data and Methods

Data Source and Study Population

This retrospective study utilized two large-scale ICU databases: the MIMIC-IV v3.1 and the eICU Collaborative Research Database. MIMIC-IV is a single-center database containing de-identified health data associated with approximately 40,000 patients who stayed in ICUs at the Beth Israel Deaconess Medical Center between 2008 and 2019. The eICU database is a multi-center database comprising data from over 200,000 ICU patients admitted to 335 units at 208 hospitals across the United States between 2014 and 2015.

The study cohort comprised adult patients (>18 years) with a diagnosis of HF who underwent echo-evaluation during their ICU stay. HF was identified using International Classification of Diseases, 9th Revision (ICD-9) code 428.0 in MIMIC-IV and corresponding ICD-9 or ICD-10 codes (428.0 or I50.9) in eICU. Echo evaluations were identified through procedural codes and treatment records. To ensure data quality and clinical relevance, we applied the following exclusion criteria: (1) non-first ICU admissions, (2) patients aged 18 years or younger, and (3) ICU length of stay less than 24 hours. The research design is shown in Fig. 1.

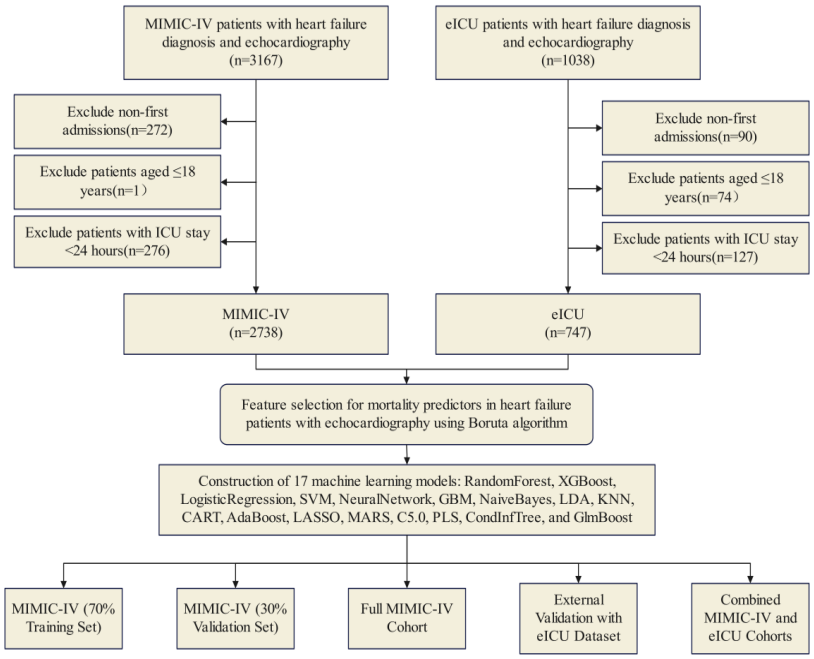

Fig. 1 Study design and patient selection flowchart

Data Extraction and Preprocessing

Data were extracted using structured query language (SQL) queries through PostgreSQL database connections. For MIMIC-IV, we first identified patients with HF diagnoses and echocardiographic assessments, then joined relevant tables to collect demographic information, clinical measurements, laboratory values, and outcomes. A similar approach was applied to the eICU database. The primary outcome was hospital mortality.

Extracted variables included demographic characteristics (age, gender, race), anthropometric measurements (height, weight, BMI), severity scores (SAPS II, OASIS, SOFA, GCS, Charlson comorbidity index), clinical interventions (mechanical ventilation, renal replacement therapy, and vasopressor use), comorbidities (atrial fibrillation, diabetes mellitus, renal disease, liver disease, chronic obstructive pulmonary disease, coronary artery disease, stroke, malignancy), vital signs (MAP, heart rate, temperature, respiratory rate, oxygen saturation), and laboratory values (complete blood count, electrolytes, renal and liver function tests, blood gases, lactate).

Missing values were addressed using multiple imputation by chained equations. For each variable type, appropriate imputation methods were applied: predictive mean matching for continuous variables, logistic regression (LR) for binary variables, and multinomial LR for categorical variables with more than two levels. Five imputed datasets were generated, and the results were aggregated by taking the mean for continuous variables and the mode for categorical variables. Variables with more than 20% missing values were excluded from the analysis.

Feature Selection

To identify the most relevant predictors of hospital mortality, we employed the Boruta algorithm, a RF-based feature selection method. Boruta works by comparing the importance of original attributes with that of their randomly permuted copies (shadow features) and progressively eliminating irrelevant features. This approach helps identify all features that are statistically significant for prediction, rather than selecting a minimal optimal subset.

The algorithm was run with 100 iterations and a significance level of 0.01. Features were categorized as "Confirmed" (significantly better than random), "Tentative" (possibly relevant), or "Rejected" (not significantly better than random). For our main analysis, we retained features confirmed as important by the Boruta algorithm to reduce dimensionality and mitigate the risk of overfitting.

Machine Learning Models

We implemented and compared 17 different ML algorithms to predict hospital mortality: Random Forest (RF); Extreme Gradient Boosting (XGBoost); Logistic Regression (LR); Support Vector Machine with Radial Basis Function Kernel (SVM); Neural Network (NN); Gradient Boosting Machine (GBM); Naive Bayes (NB); Linear Discriminant Analysis (LDA); K-Nearest Neighbors (KNN); Classification and Regression Trees (CART); Adaptive Boosting (AdaBoost); Least Absolute Shrinkage and Selection Operator (LASSO); Multivariate Adaptive Regression Splines (MARS); C5.0 Decision Tree (C5.0); Partial Least Squares (PLS); Conditional Inference Trees (CIT); Generalized Linear Model Boosting (GLMBoost).

All models were trained using the caret package in R with repeated 5-fold cross-validation (repeated 3 times) to optimize hyperparameters. The area under the receiver operating characteristic curve (AUC) was used as the primary metric for hyperparameter tuning. The MIMIC-IV dataset was split into training (70%) and internal validation (30%) sets, while the entire eICU dataset was reserved for external validation.

Model Evaluation

The performance of each model was evaluated using several metrics: Discrimination: AUC with 95% confidence intervals; Calibration: Calibration plots comparing predicted probabilities with observed outcomes; Classification metrics: Sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Models were evaluated on five different datasets: MIMIC-IV training set, MIMIC-IV testing set, entire MIMIC-IV cohort, eICU external validation set, and combined MIMIC-IV and eICU dataset. We also performed decision curve analysis (DCA) to assess the clinical utility of each model across a range of threshold probabilities.

Model Interpretability Analysis

To understand the key factors driving mortality predictions, we applied the SHAP method to the best-performing model based on overall AUC. SHAP values provide a unified measure of feature importance based on game-theoretic concepts and offer several advantages over traditional feature importance metrics:They are consistent and locally accurate; They satisfy the missingness property (features with no impact have zero SHAP values); They provide both global feature importance and local explanations for individual predictions. We calculated SHAP values using the kernelshap package in R, which implements kernel SHAP, a model-agnostic approach that can be applied to any ML algorithm. The SHAP analysis was performed separately on each dataset (MIMIC-IV training, MIMIC-IV testing, entire MIMIC-IV, eICU, and combined data) to assess the SHAP values of features and consistency of feature importance across different cohorts. SHAP dependency plots: Visualizing the relationship between feature values and their impact on predictions; SHAP beeswarm plots: Displaying the distribution of SHAP values for each feature; Waterfall plots: Showing how each feature contributes to individual predictions.

Statistical Analysis

Baseline characteristics were compared between databases and between survival and non-survival groups using appropriate statistical tests. Continuous variables were presented as median (interquartile range) and compared using the Mann-Whitney U test due to non-normal distributions. Categorical variables were presented as counts (percentages) and compared using the chi-square test or Fisher's exact test as appropriate. A two-sided p-value <0.05 was considered statistically significant. All analyses were performed using R version 4.4.2 with the following packages: caret, randomForest, XGBoost, e1071, nnet, gbm, klaR, rpart, kernelshap, shapviz, pROC, DALEX, and Boruta.

Results

Study Cohort and Baseline Characteristics

This study screened adult HF patients who underwent echo-evaluation from two ICU databases (MIMIC-IV and eICU), ultimately enrolling 3,485 patients: 2,738 from MIMIC-IV and 747 from eICU (Fig. 1). Baseline characteristics were compared between the databases (Table 1). The overall cohort had a median age of 73 years (IQR: 62-82), median weight of 80.30 kg (IQR: 67.00-96.93), with males slightly predominant (55.9%). Caucasians constituted the majority (69.8%), while Asians represented the smallest group (2.2%). Compared to eICU, MIMIC-IV patients were older (73 vs. 70 years, $p < 0.001$), lighter (79.51 vs. 85.00 kg, $p < 0.001$), and exhibited higher severity scores (SAPS II, OASIS, Charlson Comorbidity Index, all $p < 0.001$). MIMIC-IV patients also showed higher compliance rate of Sepsis 3.0 criteria (67.2% vs. 49.5%, $p < 0.001$) and vasopressor usage (43.5% vs. 15.3%, $p < 0.001$), but lower mechanical ventilation rates (47.0% vs. 54.5%, $p < 0.001$). No significant differences were observed in gender distribution, ethnicity, or laboratory parameters (sodium, potassium, creatinine; $p > 0.05$). The hospital mortality in MIMIC-IV Database and eICU Database was shown in Supplementary Table S1 and Supplementary Table S2.

Table 1. Comparison of Baseline Characteristics Between MIMIC-IV and eICU Databases

Characteristics	Level	Overall (n=3485)	eICU (n=747)	MIMIC-IV (n=2738)	P-value	SMD
Age (years) (M [Q1,Q3])		73.00 [62.00, 82.00]	70.00 [60.00, 79.00]	73.00 [63.00, 82.00]	<0.001	0.270
Gender	Female	1538 (44.1)	330 (44.2)	1208 (44.1)	1.000	0.001
	Male	1947 (55.9)	417 (55.8)	1530 (55.9)		
Race	Asian	77 (2.2)	9 (1.2)	68 (2.5)	0.001	0.171
	Black	397 (11.4)	113 (15.1)	284 (10.4)		
	Other	579 (16.6)	113 (15.1)	466 (17.0)		
	White	2432 (69.8)	512 (68.5)	1920 (70.1)		
Weight (kg) (M [Q1,Q3])		80.30 [67.00, 96.93]	85.00 [70.45, 104.30]	79.51 [66.41, 95.10]	<0.001	0.258
SAPS II score (M [Q1,Q3])		39.00 [31.00, 49.00]	32.00 [24.00, 43.00]	41.00 [33.00, 50.00]	<0.001	0.509
OASIS score (M [Q1,Q3])		33.00 [26.00, 39.00]	26.00 [21.00, 34.00]	34.00 [28.00, 40.00]	<0.001	0.683
SOFA score (M [Q1,Q3])		5.00 [3.00, 8.00]	6.00 [4.00, 8.00]	5.00 [3.00, 8.00]	0.021	0.067
GCS score (M [Q1,Q3])		15.00 [13.00, 15.00]	14.40 [11.00, 15.00]	15.00 [14.00, 15.00]	<0.001	0.296
Charlson comorbidity index (M [Q1,Q3])		6.00 [5.00, 8.00]	5.00 [3.00, 7.00]	7.00 [5.00, 8.00]	<0.001	0.726
Sepsis-3 criteria met	No	1275 (36.6)	377 (50.5)	898 (32.8)	<0.001	0.364

Characteristics	Level	Overall (n=3485)	eICU (n=747)	MIMIC-IV (n=2738)	P-value	SMD
	Yes	2210 (63.4)	370 (49.5)	1840 (67.2)		
Mechanical ventilation	No	1792 (51.4)	340 (45.5)	1452 (53.0)	<0.001	0.151
	Yes	1693 (48.6)	407 (54.5)	1286 (47.0)		
Renal replacement therapy	No	3313 (95.1)	704 (94.2)	2609 (95.3)	0.283	0.047
	Yes	172 (4.9)	43 (5.8)	129 (4.7)		
Vasopressor use	No	2180 (62.6)	633 (84.7)	1547 (56.5)	<0.001	0.652
	Yes	1305 (37.4)	114 (15.3)	1191 (43.5)		
Sedative use	No	2096 (60.1)	609 (81.5)	1487 (54.3)	<0.001	0.610
	Yes	1389 (39.9)	138 (18.5)	1251 (45.7)		
Heart failure	Yes	3485 (100.0)	747 (100.0)	2738 (100.0)	NA	<0.001
Atrial fibrillation	No	1721 (49.4)	384 (51.4)	1337 (48.8)	0.228	0.052
	Yes	1764 (50.6)	363 (48.6)	1401 (51.2)		
Diabetes mellitus	No	2077 (59.6)	403 (53.9)	1674 (61.1)	<0.001	0.146
	Yes	1408 (40.4)	344 (46.1)	1064 (38.9)		
Renal disease	No	2231 (64.0)	523 (70.0)	1708 (62.4)	<0.001	0.162
	Yes	1254 (36.0)	224 (30.0)	1030 (37.6)		
Liver disease	No	3267 (93.7)	715 (95.7)	2552 (93.2)	0.015	0.110
	Yes	218 (6.3)	32 (4.3)	186 (6.8)		
COPD	No	2642 (75.8)	475 (63.6)	2167 (79.1)	<0.001	0.349
	Yes	843 (24.2)	272 (36.4)	571 (20.9)		
Coronary artery disease	No	1602 (46.0)	185 (24.8)	1417 (51.8)	<0.001	0.578
	Yes	1883 (54.0)	562 (75.2)	1321 (48.2)		
Stroke	No	3117 (89.4)	638 (85.4)	2479 (90.5)	<0.001	0.158
	Yes	368 (10.6)	109 (14.6)	259 (9.5)		
Malignancy	No	2985 (85.7)	640 (85.7)	2345 (85.6)	1.000	0.001
	Yes	500 (14.3)	107 (14.3)	393 (14.4)		
MAP (mmHg) (M [Q1,Q3])		79.00 [67.00, 92.00]	83.00 [70.00, 98.00]	78.00 [67.00, 90.00]	<0.001	0.273
Hr (bpm) (M [Q1,Q3])		88.00 [75.00, 104.00]	89.00 [75.50, 105.00]	88.00 [75.00, 103.00]	0.940	0.007
Temperature (°C) (M [Q1,Q3])		36.61 [36.22, 37.00]	36.70 [36.39, 37.00]	36.56 [36.17, 37.00]	<0.001	0.114
Rr (breaths/min) (M [Q1,Q3])		20.00 [16.00, 24.00]	20.00 [17.00, 25.00]	20.00 [16.00, 24.00]	<0.001	0.175
SpO_2 (M [Q1,Q3])		97.00 [94.00, 100.00]	97.00 [94.00, 99.00]	97.00 [94.00, 100.00]	0.068	0.061
WBC count (×10⁹/L) (M [Q1,Q3])		10.80 [7.80, 14.80]	10.30 [7.70, 14.05]	10.90 [7.80, 15.10]	0.036	0.082
Hb (g/dL) (M [Q1,Q3])		10.40 [9.00, 12.00]	11.00 [9.60, 12.30]	10.23 [8.90, 11.80]	<0.001	0.298
PLT (×10⁹/L) (M [Q1,Q3])		205.00 [148.00, 271.00]	197.00 [148.00, 255.90]	208.00 [148.25, 276.75]	0.004	0.169
Sodium (mmol/L) (M [Q1,Q3])		138.00 [135.00,	138.00 [135.00,	138.00 [134.00,	0.392	0.057

Characteristics	Level	Overall (n=3485)	eICU (n=747)	MIMIC-IV (n=2738)	P-value	SMD
		141.00]	141.00]	141.00]		
Potassium (mmol/L) (M [Q1,Q3])		4.20 [3.80, 4.70]	4.10 [3.70, 4.60]	4.20 [3.80, 4.70]	0.002	0.135
Bicarbonate (mmol/L) (M [Q1,Q3])		24.00 [21.00, 28.00]	25.10 [22.00, 29.00]	24.00 [21.00, 27.00]	<0.001	0.330
Chloride (mmol/L) (M [Q1,Q3])		103.00 [99.00, 107.00]	102.00 [98.00, 106.00]	103.00 [99.00, 107.00]	<0.001	0.145
BUN (mg/dL) (M [Q1,Q3])		29.00 [19.00, 45.00]	27.00 [19.00, 43.00]	29.00 [19.00, 46.00]	0.032	0.090
Creatinine (mg/dL) (M [Q1,Q3])		1.30 [0.90, 2.05]	1.31 [0.92, 2.02]	1.30 [0.90, 2.10]	0.613	0.015

Note: median, M ;interquartile range, IQR; Quartile 1, Q1; Quartile 3, Q3; chronic obstructive pulmonary disease, COPD; mean arterial pressure, MAP;Heart rate, HR; Respiratory rate, Rr; white blood cell, WBC; Hemoglobin, Hb; Platelet count, PLT; blood urea nitrogen, BUN.

Feature Selection and Correlation Analysis

The Boruta algorithm was employed to screen initial features, identifying 32 "confirmed" features (Fig. 2A). The feature importance rankings indicated that the SAPS II score was the top contributor, followed by OASIS score, SOFA score, first 24-hour BUN, and GCS. Other significant features included first 24-hour creatinine, bicarbonate (HCO_3^-), systolic blood pressure (SBP), and MAP. Additionally, the Charlson Comorbidity Index, age, body temperature, WBC, and vasoactive medication use (e.g., vasopressin, norepinephrine) were confirmed as important predictors. In the MIMIC-IV dataset (Fig. 2B), correlation analysis among confirmed important features revealed the strongest correlation between MAP and diastolic blood pressure (DBP) ($r = 0.866$, $p < 0.001$), followed by mechanical ventilation and sedative use ($r = 0.834$, $p < 0.001$) and midazolam-fentanyl co-administration ($r = 0.760$, $p < 0.001$). These results revealed significant associations between clinical interventions and vital signs. In the eICU dataset (Fig. 2C), correlation analysis similarly identified MAP-DBP as the strongest correlation ($r = 0.894$, $p < 0.001$), followed by MAP-SBP ($r = 0.773$, $p < 0.001$). Furthermore, SAPS II and OASIS scores exhibited strong correlation ($r = 0.765$, $p < 0.001$), indicating consistency across disease severity scoring systems.

ISSN: 0378-1844

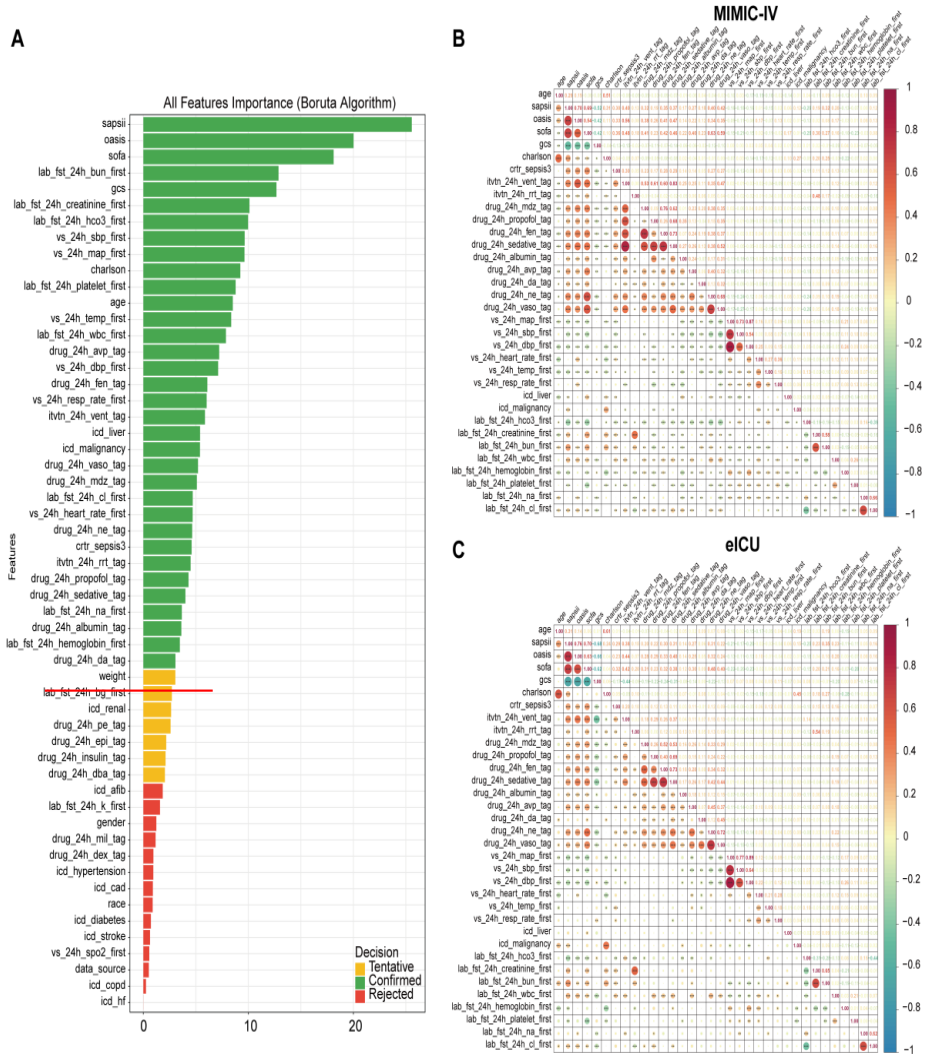

Fig. 2 Feature selection and correlation analysis.

(A) Boruta algorithm analysis identifying confirmed and tentative important features for predicting mortality in CHF patients with echocardiographic evaluation. (B) Correlation matrix of confirmed features in the MIMIC-IV dataset. The lower triangle displays correlation coefficients with significance levels (*p < 0.05, **p < 0.01, ***p < 0.001), while the upper triangle shows numerical correlation values. (C) Correlation matrix of confirmed features in the eICU dataset. The lower triangle displays correlation coefficients with significance levels (*p < 0.05, **p < 0.01, ***p < 0.001), while the upper triangle shows numerical correlation values.

Model performance evaluation

We compared 17 ML models for predicting in-hospital mortality among ICU patients. All models showed acceptable predictive capabilities, with the RF model exhibiting optimal performance. In the MIMIC-IV training set, the RF model achieved an AUC of 1.000 (Fig. 3A). In the MIMIC-IV test set, both RF and GBM showed the highest AUC of 0.778 (Fig. 3B). For the external validation set (eICU), the LR model performed best with an AUC of 0.782 (Fig. 3C), demonstrating robust generalization capability. When evaluated on the entire MIMIC-IV dataset, RF again demonstrated exceptional performance (AUC = 0.969; Fig. 3D), confirming its effectiveness with large-scale single-center data. Cross-dataset integrated analysis (MIMIC + eICU) ultimately revealed RF's superior stability and performance (highest AUC = 0.940; Fig. 3E). These results strongly support our proposed prediction model and establish RF as the benchmark for in-hospital mortality prediction.

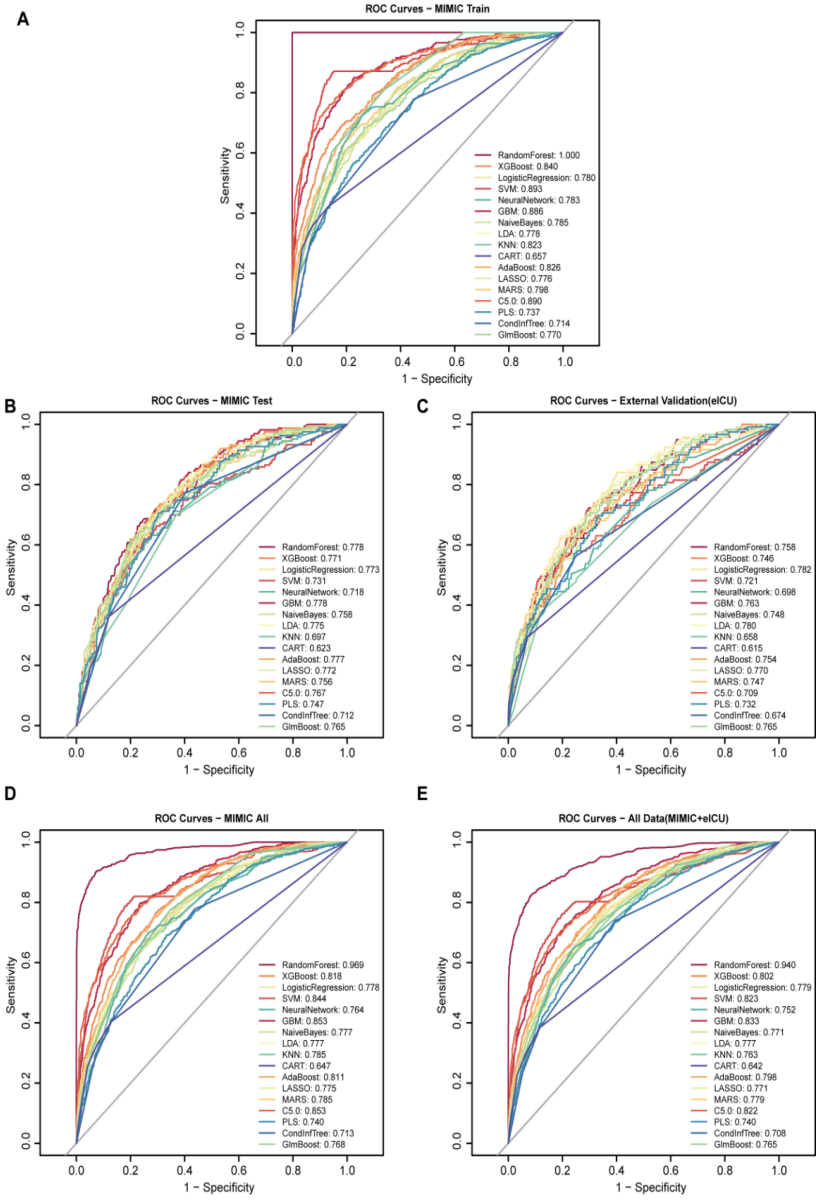

Fig. 3 ROC curves for all machine learning models across different datasets.
(A) ROC curves for the MIMIC-IV training cohort. (B) ROC curves for the MIMIC-IV testing cohort (internal validation). (C) ROC curves for the eICU cohort (external validation). (D) ROC curves for the entire MIMIC-IV dataset. (E) ROC curves for the combined MIMIC-IV and eICU datasets.

Model Interpretability Analysis

This study employed the SHAP method to systematically evaluate the global feature importance of the model. The contribution ranking of key features across multiple datasets (Fig. 4A) showed that SAPS II score, OASIS score, and SOFA score ranked in the top three, indicating the core role of disease severity scores in predicting in-hospital mortality. Additionally, age, first 24-hour WBC, and first 24-hour BUN values also demonstrated significant contributions, further validating the importance of these features in clinical practice. Notably, the high consistency in the ranking of these key features across different datasets highlights the stability and reliability of the model.

Through SHAP dependence plots (Fig. 4B), an in-depth analysis was conducted on the relationship between major mortality predictors and model outputs. The results showed that when the SAPS II score was below 45, the SHAP value was negative; after exceeding 45, the mortality risk rapidly increased with higher scores, and the rate of increase slowed down after 60. The relationship between SOFA score and mortality risk exhibited a similar trend: when the score exceeded 8, mortality risk was positively correlated with the score; however, when the score surpassed 15, the mortality risk decreased. For the OASIS score, an increase in the score was negatively correlated with mortality risk below 20; between 20 and 40, SHAP values remained negative; after exceeding 40, mortality risk gradually rose, peaked around 50, followed by a slight decline, and rose again after 60. As a predictor, age showed a significant increase in mortality risk for patients over 70 years old. The trend of first 24-hour WBC first decreased and then increased, peaking at 25 × 10^9/L before gradually declining. Higher first 24-hour BUN values were overall associated with increased mortality risk, further indicating a high consistency between the clinical significance of these features and the model's predictions.

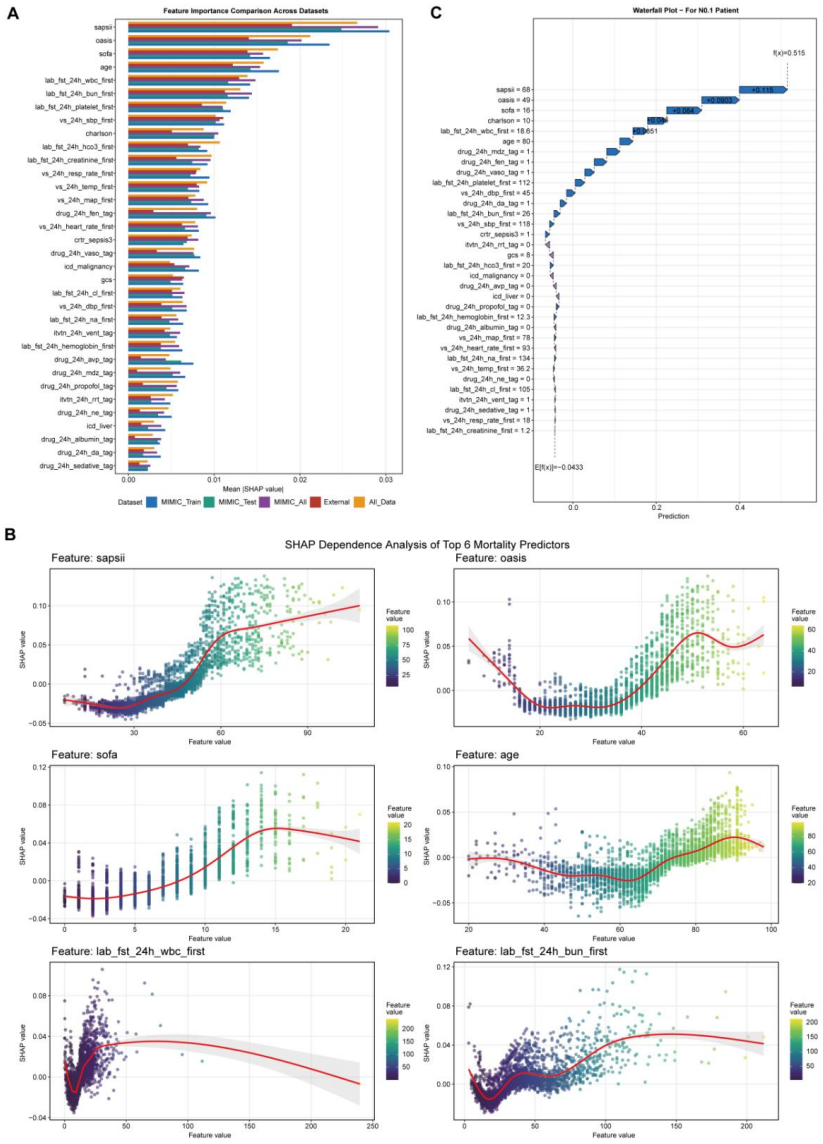

Fig. 4 Model interpretability analysis using SHAP.

(A) Comparative feature importance across different datasets, with features ranked by mean absolute SHAP value. Bars represent the contribution of each feature to the model prediction across datasets. (B) SHAP dependence plots for the top six mortality predictors, illustrating the relationship between feature values (x-axis) and their impact on model output (y-axis). Higher SHAP values indicate increased mortality risk. (C) Waterfall plot for a representative patient case, demonstrating how each feature contributes to the final prediction.

To further enhance the clinical interpretability of the model, this study used a waterfall plot to illustrate the prediction process for a representative patient, detailing the contributions of each feature to the final prediction outcome (Fig. 4C). Taking an 80-year-old patient as an example, with SAPS II score of 68, OASIS score of 49, SOFA score of 16, Charlson Comorbidity Index of 10, WBC of 18.6 × 10^9/L, and blood urea nitrogen of 26mg/dL, the positive contributions (SHAP values) of these features were 0.115, 0.090, 0.084, 0.046, 0.035, and 0.015, respectively, collectively elevating the predicted mortality risk for this patient. Other features such as mechanical ventilation and vasoactive medication use also provided varying degrees of contribution, demonstrating the model's granularity and transparency in individualized prediction.

Discussion

In this study, we compared 17 ML algorithms across multiple databases and constructed a mortality prediction model for HF patients based on echo-evaluation using the optimal algorithm, with the RF model demonstrating the best predictive performance. Results from the external validation cohort further confirmed the model' s stability and accuracy. By applying the SHAP algorithm to interpret the RF model, we identified the core roles of SAPS II score, OASIS score, SOFA score, age, WBC, and first 24-hour BUN values in predicting in-hospital mortality. The feature variables incorporated into the model are clinically accessible and easy to implement in primary healthcare settings.

The RF algorithm, first proposed by Breiman[26] in 2001, is a ML method that utilizes the Bootstrap method for sample training and prediction. It calculates the impact of each variable on node observations in classification trees using the Gini index and possesses feature importance evaluation capabilities. The primary advantage of RF lies in its ability to directly estimate generalization error through out-of-bag (OOB) error without requiring separate test set partitioning, demonstrating strong anti-overfitting capabilities and generalization performance[27]. In this study, the RF model exhibited optimal predictive performance across internal and external datasets, with combined dataset analysis revealing its highest AUC of 0.940. Similarly, Angraal *et al*.[28] compared five ML algorithms and found RF to perform best in predicting mortality among HF patients with preserved ejection fraction. Chen *et al*.[29] and Masetic *et al*.[30] also emphasized the critical role of RF in HF risk stratification and prognosis. These results collectively highlight the undeniable clinical value of RF in HF prognosis management. In summary, current research on HF prognostic models predominantly focuses on comparing single ML models with traditional approaches, with limited studies comparing multiple ML algorithms, particularly lacking reports targeting ICU-managed HF patients. Additionally, most studies lack external validation, relying solely on single-source data, thereby limiting

clinical generalizability[31-34]. Therefore, this study comparatively analyzed ML models across multiple data sources and selected the optimal RF model, which holds high clinical value for mortality prediction and provides a reference for selecting the most suitable predictive model in clinical practice.

Studies indicate that HF patients often develop multi-organ failure as their condition worsens, exacerbating poor prognosis[9, 35]. Severity scoring systems such as SAPS II, OASIS, and SOFA dynamically monitor multi-organ functional status, comprehensively assess the condition of cardiac patients, and have become critical tools for prognosis prediction. Notably, these three scoring systems play pivotal roles in ICU patient prognosis management. Research shows that SAPS II stably predicts long-term ICU patient outcomes (AUC: 0.7-0.8), while OASIS and SOFA demonstrate unique advantages in rapid mortality risk assessment and dynamic organ function monitoring[36-39]. In this study, preliminary Boruta algorithm evaluation and cross-dataset SHAP analysis consistently ranked these three scoring systems highly, confirming their status as key variables for predicting ICU HF mortality. Additionally, age, first 24-hour WBC, and first 24-hour BUN values ranked prominently among predictors. Existing studies demonstrate that HF incidence and mortality correlate positively with age[40], particularly showing exponential risk escalation after 60 years[41], aligning with our predictions. Meanwhile, WBC levels, established as independent predictors of HF progression and mortality, hold high weight in prognostic models, potentially due to their association with systemic inflammation activation and pathophysiological processes[42, 43]. Similarly, Li *et al*.[31] and Mpanya *et al*.[33] emphasized BUN's critical predictive role. BUN's importance in ML-based HF mortality prediction[44], may stem from its function as a renal biomarker reflecting protein metabolism; elevated BUN levels correlate with HF severity[45]. In conclusion, this study identified six key risk predictors (SAPS II, OASIS, SOFA, age, WBC, and BUN) that play decisive roles in HF mortality prediction, providing a basis for clinical identification of high-risk HF patients.

Building on these findings, we further analyzed these risk predictors' contributions to the RF model. Notably, ML models have long been viewed as "black boxes" due to their complex and opaque decision-making processes. Lundberg and Lee proposed the SHAP framework to interpret such models and quantify feature importance[46]. In this study, SHAP values were applied to interpret the RF model's predictions, enhancing its clinical interpretability. Among mortality predictors, severity scores played central roles, with SAPS II having the highest SHAP value. Mortality risk escalated rapidly with increasing SAPS II scores, reflecting multi-organ dysfunction's critical impact on HF prognosis. Interestingly, mortality risk increase decelerated when SAPS II exceeded a threshold of 60. This may occur because critically ill patients reach physiological compensation limits for multi-organ failure, reducing the marginal contribution of

additional pathological variables to overall risk. Additionally, SAPS II's lower sensitivity to dynamic changes compared to SOFA and OASIS might explain its flattened SHAP slope in high-score ranges[36]. Similarly, SOFA exhibited a comparable trend to SAPS II, but mortality risk decreased when scores surpassed 15, possibly due to interactions with clinical interventions (e.g., treatments weakening its predictive power). Chen *et al*.[47] observed reduced SOFA predictive efficacy when combined with lactate/vasopressor metrics. Intriguingly, OASIS showed more fluctuating mortality risk associations than SAPS II/SOFA. When the OASIS score is below 20, pain symptoms may systematically reduce its mortality prediction effect[48], while scores above 40 showed elevated SHAP contributions due to interactions between respiratory/blood pressure parameters and hemoglobin levels[49]. Meanwhile, the observed fluctuations in OASIS score SHAP values beyond 50 in this study may be attributed to the OASIS scoring system's reliance on linear assumptions inherent to LR. When nonlinear data or limited high-score samples are present, residuals may exhibit abnormal volatility, thereby diminishing its predictive efficacy[50]. Furthermore, SHAP analysis also highlighted age, WBC, and BUN's roles in HF mortality prediction. Age and BUN showed positive correlations with mortality risk, aligning with clinical expectations. However, WBC's contribution peaked at 25 × 10^9/L then declined, potentially reflecting post-acute phase inflammation resolution or treatment effects. Finally, a waterfall plot analysis of a representative patient's feature contributions demonstrated uniformly positive SHAP values, reaffirming the model's clinical utility in individualized prediction.

This study demonstrates significant innovation and clinical applicability in constructing a prognostic prediction model for HF. Compared to previous studies relying predominantly on single-dimensional clinical indicators, we systematically integrated echo parameters with multi-dimensional clinical features, like demographic characteristics, vital signs, laboratory findings, and clinical interventions and so on. This multi-modal data integration strategy not only enhances the model’s biological plausibility but also addresses the limitations of missing imaging parameters in traditional studies. Methodologically, we ensured model generalizability and reliability by comparing 17 ML algorithms and conducting internal/external validation across dual databases (MIMIC+eICU), effectively mitigating bias due to algorithmic preference. Unlike most prognostic models targeting general ward HF patients, this study specifically developed a model for HF patients in the ICU setting. ICU patients exhibit more complex pathophysiological changes and greater susceptibility to intervention impacts, and our model better adapts to these critical care complexities, significantly improving clinical utility through precise population targeting. Furthermore, SHAP analysis enhanced model interpretability. Compared to traditional scoring systems, our model incorporates dynamic variables reflecting disease severity and interventions,

overcoming the static limitations of single scoring systems and aligning with ICU dynamic management needs. Through this approach, we not only provide accurate predictions but also identify key prognostic determinants and their interaction patterns, offering reliable evidence for clinical decision-making.

However, this study has certain limitations. The limited clinical data and sample size, coupled with underrepresentation of Asian populations, may restrict the generalizability of our findings to Asian populations. Although multi-database analysis was employed, variations in data quality and completeness may exist. Furthermore, ICU admission decisions might involve non-clinical factors such as hospital capacity and institutional policies, potentially introducing bias into our predictions. Therefore, we plan to include additional cohorts in future studies to further validate the model's effectiveness. Second, Echo inherently carries operator-dependent subjectivity, with potential interpretation variability among examiners. Advanced imaging data such as CMR could provide more reliable details on cardiac structure and function. Despite the current challenges of limited number of patients undergoing CMR and significant data extraction difficulties, we aim to incorporate such advanced imaging modalities in future work to enhance prediction accuracy, considering their potential value as technology advances. Additionally, while we used the eICU dataset for external validation, the applicability of our findings to other populations remains uncertain, necessitating further prospective studies to verify the model's accuracy. Through ongoing research, we aim to develop more precise, reliable, and broadly applicable HF prognostic models to improve clinical diagnosis and treatment. This effort will not only facilitate early identification of high-risk patients but also provide scientific evidence for personalized intervention strategies, ultimately improving patient outcomes and reducing healthcare burdens.

Conclusion

In summary, this study integrates multimodal data, including echo diagnostic findings, clinical severity scores, laboratory indices, and demographic characteristics, and conducted a comparative analysis of 17 machine learning models to developed a prediction model for in-hospital mortality among ICU-managed HF patients. Among these algorithms, the RF model demonstrates exceptional predictive performance. This systematic comparison provides critical reference for selecting the most appropriate predictive model in clinical practice. Additionally, SHAP enhances model interpretability, and the incorporated features combine clinical accessibility with generalizability, offering actionable references for clinical decision-making and personalized precision treatment of HF patients in the ICU.

Abbreviations

HF	Heart Failure
ICU	Intensive Care Unit
ML	Machine Learning
RF	Random Forest
ROC	Receiver Operating Characteristic
AUC	Area Under The Receiver Operating Characteristic
SHAP	Shapley Additive Explanations
WBC	White Blood Cell Count
BUN	Blood Urea Nitrogen
CHF	Congestive Heart Failure
Echo	Echocardiography
AI	Artificial Intelligence
SQL	Structured Query Language
BMI	Body Mass Index
MICE	Multiple Imputation By Chained Equations
XGBoost	Extreme Gradient Boosting
LR	Logistic Regression
SVM	Support Vector Machine With Radial Basis Function Kernel
NN	Neural Network
GBM	Gradient Boosting Machine

NB	Naive Bayes
LDA	Linear Discriminant Analysis
KNN	K-Nearest Neighbor
CART	Classification And Regression Trees
AdaBoost	Adaptive Boosting
LASSO	Least Absolute Shrinkage And Selection Operator
MARS	Multivariate Adaptive Regression Splines
C5.0	C5.0 Decision Tree
PLS	Partial Least Squares
CIT	Conditional Inference Trees
GLMBoost	Generalized Linear Model Boosting
PPV	Positive Predictive Value
NPV	Negative Predictive Value
DCA	Decision Curve Analysis
GCS	Glasgow Coma Scale
SBP	Systolic Blood Pressure
COPD	Chronic Obstructive Pulmonary Disease
MAP	Mean Arterial Pressure
DBP	Diastolic Blood Pressure
OOB	Out-Of-Bag
CMR	Cardiac Magnetic Resonance Imaging

Declarations

Ethics approval and consent to participate

This study was conducted using de-identified patient data from MIMIC-IV and eICU databases. The use of MIMIC-IV was approved by the Institutional Review Board of Beth Israel Deaconess Medical Center (Boston, MA, USA) and Massachusetts Institute of Technology (Cambridge, MA, USA), with waiver of informed consent. The eICU database was also certified as de-identified, and its usage was approved under project number 67424917.

Consent for publication

Agree to publish.

Availability of data and materials

The data that support the findings of this study are derived from MIMIC-IV (version 3.1) and eICU Collaborative Research Database, which are publicly available with appropriate credentialing. MIMIC-IV is available through PhysioNet (https://physionet.org/content/mimiciv/3.1/) after completion of required training and approval. The eICU Collaborative Research Database is available through the eICU Collaborative Research Database project page (https://eicu-crd.mit.edu/) after registration and approval. Analysis code is available from the corresponding author upon reasonable request.

Competing interests

The authors declare no conflicts of interest applicable to this study.

Funding

No funding.

Authors' contributions

Dan Xu:Conceptualization, Formal analysis, Investigation, Writing original draft; Hongli Liu:Resources, Software, Writing original draft; Jing Yang: Methodolog, Supervision, Visualization, Writing original draft, Writing review and editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We would like to thank the teams responsible for the development and maintenance of the MIMIC-IV and eICU Collaborative Research Database, which have made this research possible.

References

1. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS: **Global burden of heart failure: a comprehensive and updated review of epidemiology.** *Cardiovasc Res* 2023, **118:**3272-3287.
2. Cook C, Cole G, Asaria P, Jabbour R, Francis DP: **The annual global economic burden of heart failure.** *Int J Cardiol* 2014, **171:**368-376.
3. Roger VL: **Epidemiology of Heart Failure: A Contemporary Perspective.** *Circ Res* 2021, **128:**1421-1434.
4. Roth GA, Dorsey H, Decleene N, Razo C, Stark B, Johnson C, Collaborators TGBoDS: **The global burden of heart failure: a systematic analysis for the Global Burden of Disease Study 2021.** *European Heart Journal* 2023, **44**.
5. Díez-Villanueva P, Jiménez-Méndez C, Alfonso F: **Heart failure in the elderly.** *J Geriatr Cardiol* 2021, **18:**219-232.
6. Wunsch H, Angus DC, Harrison DA, Collange O, Fowler R, Hoste EA, de Keizer NF, Kersten A, Linde-Zwirble WT, Sandiumenge A, Rowan KM: **Variation in critical care services across North America and Western Europe.** *Crit Care Med* 2008, **36:**2787-2793, e2781-2789.
7. Adams KF, Jr., Fonarow GC, Emerman CL, LeJemtel TH, Costanzo MR, Abraham WT, Berkowitz RL, Galvao M, Horton DP: **Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE).** *Am Heart J* 2005, **149:**209-216.
8. Abdin A, Anker SD, Butler J, Coats AJS, Kindermann I, Lainscak M, Lund LH, Metra M, Mullens W, Rosano G, et al: **'Time is prognosis' in heart failure: time-to-treatment initiation as a modifiable risk factor.** *ESC Heart Fail* 2021, **8:**4444-4453.
9. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, Burri H, Butler J, Čelutkienė J, Chioncel O, et al: **2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure.** *Eur Heart J* 2021, **42:**3599-3726.
10. Martini L, Lisi M, Pastore MC, Righini FM, Rubboli A, Henein MY, Cameli M: **The Role of Speckle Tracking Echocardiography in the Evaluation of Advanced-Heart-Failure Patients.** *J Clin Med* 2024, **13**.
11. Flachskampf F, Baron T: **Heart failure and cardiac imaging: Choosing wisely in the era of multimodality imaging.** *Anatol J Cardiol* 2020, **23:**204-208.

12. Tanisha, Amudha C, Raake M, Samuel D, Aggarwal S, Bashir ZMD, Marole KK, Maryam I, Nazir Z: **Diagnostic Modalities in Heart Failure: A Narrative Review.** *Cureus* 2024, **16:**e67432.
13. Jia H, Liao S, Zhu X, Liu W, Xu Y, Ge R, Zhu Y: **Deep learning prediction of survival in patients with heart failure using chest radiographs.** *Int J Cardiovasc Imaging* 2024, **40:**1891-1901.
14. Screever EM, van der Wal MHL, van Veldhuisen DJ, Jaarsma T, Koops A, van Dijk KS, Warink-Riemersma J, Coster JE, Westenbrink BD, van der Meer P, et al: **Comorbidities complicating heart failure: changes over the last 15 years.** *Clin Res Cardiol* 2023, **112:**123-133.
15. Kaur D, Hughes JW, Rogers AJ, Kang G, Narayan SM, Ashley EA, Perez MV: **Race, Sex, and Age Disparities in the Performance of ECG Deep Learning Models Predicting Heart Failure.** *Circ Heart Fail* 2024, **17:**e010879.
16. Alsharqi M, Upton R, Mumith A, Leeson P: **Artificial intelligence: a new clinical support tool for stress echocardiography.** *Expert Review of Medical Devices*.
17. Yoon YE, Kim S, Chang HJ: **Artificial Intelligence and Echocardiography.** *J Cardiovasc Imaging* 2021, **29:**193-204.
18. Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, Išgum I: **State-of-the-Art Deep Learning in Cardiovascular Image Analysis.** *JACC Cardiovasc Imaging* 2019, **12:**1549-1565.
19. Alvin, Rajkomar, Jeffrey, Dean, Isaac, Kohane: **Machine Learning in Medicine.** *The New England journal of medicine* 2019.
20. Zhou J, Du M, Chang S, Chen Z: **Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis.** *Cardiovascular Ultrasound* 2021, **19**.
21. Averbuch T, Sullivan K, Sauer A, Mamas MA, Voors AA, Gale CP, Metra M, Ravindra N, Van Spall HGC: **Applications of artificial intelligence and machine learning in heart failure.** *Eur Heart J Digit Health* 2022, **3:**311-322.
22. Qinyu Z, Wang H, Luo JC, Luo M, Luo Z: **Development and Validation of a Machine-Learning Model for Prediction of Extubation Failure in Intensive Care Units.** *Frontiers in Medicine* 2021, **8:**676343.
23. Zhou J, Du M, Chang S, Chen Z: **Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis.** *Cardiovasc Ultrasound* 2021, **19:**29.
24. Valsaraj A, Kalmady SV, Sharma V, Frost M, Sun W, Sepehrvand N, Ong M, Equilbec C, Dyck JRB, Anderson T, et al: **Development and validation of echocardiography-based machine-learning models to predict mortality.** *EBioMedicine* 2023, **90:**104479.

25. Ulloa Cerna AE, Jing L, Good CW, vanMaanen DP, Raghunath S, Suever JD, Nevius CD, Wehner GJ, Hartzel DN, Leader JB, et al: **Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality.** *Nat Biomed Eng* 2021, **5:**546-554.
26. Breiman: **Random forests.** *MACH LEARN* 2001, **2001,45(1):**5-32.
27. Hu J, Szymczak S: **A review on longitudinal data analysis with random forest.** *Brief Bioinform* 2023, **24**.
28. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM: **Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction.** *JACC Heart Fail* 2020, **8:**12-21.
29. Chen H, Jiang R, Huang W, Chen K, Zeng R, Wu H, Yang Q, Guo K, Li J, Wei R, et al: **Identification of energy metabolism-related biomarkers for risk prediction of heart failure patients using random forest algorithm.** *Front Cardiovasc Med* 2022, **9:**993142.
30. Masetic Z, Subasi A: **Congestive heart failure detection using random forest classifier.** *Comput Methods Programs Biomed* 2016, **130:**54-64.
31. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z: **Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database.** *BMJ Open* 2021, **11:**e044779.
32. Chen Z, Li T, Guo S, Zeng D, Wang K: **Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure.** *Frontiers in Cardiovascular Medicine* 2023, **10:**10.
33. Mpanya D, Celik T, Klug E, Ntsinjana H: **Machine learning and statistical methods for predicting mortality in heart failure.** *Heart Fail Rev* 2021, **26:**545-552.
34. Li L, Zhang Z, Xiong Y, Hu Z, Liu S, Tu B, Yao Y: **Prediction of hospital mortality in mechanically ventilated patients with congestive heart failure using machine learning approaches.** *International Journal of Cardiology* 2022**:**358.
35. Wilk MM, Wilk J, Urban S, Gajewski P: **Current Review of Heart Failure-Related Risk and Prognostic Factors.** *Biomedicines* 2024, **12**.
36. Xu F, Li W, Zhang C, Cao R: **Performance of Sequential Organ Failure Assessment and Simplified Acute Physiology Score II for Post-Cardiac Surgery Patients in Intensive Care Unit.** *Front Cardiovasc Med* 2021, **8:**774935.
37. Choi JY, Jang JH, Lim YS, Jang JY, Lee G, Yang HJ, Cho JS, Hyun SY: **Performance on the APACHE II, SAPS II, SOFA and the OHCA score of post-cardiac arrest patients treated with therapeutic hypothermia.** *PLoS One* 2018, **13:**e0196197.

38. Elias A, Agbarieh R, Saliba W, Khoury J, Bahouth F, Nashashibi J, Azzam ZS: **SOFA score and short-term mortality in acute decompensated heart failure.** *Sci Rep* 2020, **10:**20802.
39. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA: **Performance of intensive care unit severity scoring systems across different ethnicities.** *medRxiv* 2021.
40. Li T, Jiang YL, Kang J, Song S, Du QF, Yi XD: **Prevalence and risk factors of frailty in older patients with chronic heart failure: a systematic review and meta-analysis.** *Aging Clin Exp Res* 2023, **35:**2861-2871.
41. Li Y, Xu J, Liu X, Wang X, Zhao C, He K: **Development and validation of an integrated prognostic model for all-cause mortality in heart failure: a comprehensive analysis combining clinical, electrocardiographic, and echocardiographic parameters.** *BMC Cardiovasc Disord* 2025, **25:**221.
42. Bekwelem W, Lutsey PL, Loehr LR, Agarwal SK, Astor BC, Guild C, Ballantyne CM, Folsom AR: **White blood cell count, C-reactive protein, and incident heart failure in the Atherosclerosis Risk in Communities (ARIC) Study.** *Ann Epidemiol* 2011, **21:**739-748.
43. Barron HV, Cannon CP, Murphy SA, Braunwald E, Gibson CM: **Association between white blood cell count, epicardial blood flow, myocardial perfusion, and clinical outcomes in the setting of acute myocardial infarction: a thrombolysis in myocardial infarction 10 substudy.** *Circulation* 2000, **102:**2329-2334.
44. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, Zhu W, Sama I, Tadel M, Campagnari C, et al: **Improving risk prediction in heart failure using machine learning.** *Eur J Heart Fail* 2020, **22:**139-147.
45. Kazory A: **Emergence of blood urea nitrogen as a biomarker of neurohormonal activation in heart failure.** *Am J Cardiol* 2010, **106:**694-700.
46. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK, Newman SF, Kim J, Lee SI: **Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.** *Nat Biomed Eng* 2018, **2:**749-760.
47. Chen S, Fan J, Pishgar E, Alaei K, Placencia G, Pishgar M: **XGBoost-Based Prediction of ICU Mortality in Sepsis-Associated Acute Kidney Injury Patients Using MIMIC-IV Database with Validation from eICU Database.** 2025.
48. Zhang J, Bolanos Trujillo LD, Tanwar A, Ive J, Gupta V, Guo Y: **Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use.** *BMJ Health Care Inform* 2022, **29**.
49. Greenberg JD, Huberts LCE, Ritchie A, Ooi S-Y, Flynn GM, Hart GK, Gallego B: **Improved Sensitivity For Detection Of Clinical Deterioration**

When Diagnostic Pathology And Patient Trends Are Included In Machine Learning Models. *medRxiv* 2024**:**2024.2010.2020.24315403.

50. Baptista ML, Goebel K, Henriques EMP: **Relation between prognostics predictor evaluation metrics and local interpretability SHAP values.** *Artificial Intelligence* 2022.

Supplementary materials

Table S1. Comparison of Hospital Mortality in MIMIC-IV Database

Table S2. Comparison of Hospital Mortality in eICU Database