British Journal of Hospital Medicine

Br. J. Hosp. Med. (Lond) **2026**; 87(9): 55546
https://doi.org/10.31083/BJHM55546

Article

Development and Validation of an Interpretable Machine Learning–Based Prediction Model of Sepsis: A Retrospective Multicenter Cohort Study

Zhangcong Lu[1,†], Linghan Hu[2,†], Pengfei Shui[2], Binyu Zhao[3], Shuyu Zhan[2], Chen Huang[2], Jiajie Huang[2], Yang He[2], Dongpeng Li[3], Yucai Hong[2,*], Ning Liu[2,4,*]

[1]Department of Nursing, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 310016 Hangzhou, Zhejiang, China
[2]Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, 310016 Hangzhou, Zhejiang, China
[3]Department of Nursing, School of Medicine, Hangzhou City University, 310015 Hangzhou, Zhejiang, China
[4]Zhejiang Key Laboratory of Intelligent Medical Decision Support, Taizhou Institute of Zhejiang University, 318000 Taizhou, Zhejiang, China
*Correspondence: realhealth@zju.edu.cn (Yucai Hong); liuning8905@zju.edu.cn (Ning Liu)
[†]These authors contributed equally.
Academic Editor: Jessie Welbourne
Submitted: 2 July 2026 Revised: 10 August 2026 Accepted: 21 August 2026 Published: 20 September 2026

Abstract

Aims/Background: Sepsis is a leading cause of morbidity and mortality in intensive care units (ICUs). Early identification of patients at high risk of death is essential to guide timely interventions and optimize resource allocation. This study aimed to develop and validate an interpretable machine learning model for predicting 28-day mortality in adult ICU patients. **Methods:** We retrospectively analyzed data from three independent critical care databases, including 98,233 adult ICU patients. Sixteen routinely collected variables available by the 24-hour prediction landmark—including vital signs and laboratory parameters obtained during the first 24 hours after ICU admission, baseline comorbidities documented at or before ICU admission, and treatments administered during the same 24-hour period—were selected for model development. Nine machine learning algorithms were compared, and model performance was evaluated using both internal and external validation cohorts. Explainable machine learning techniques were applied to identify the most influential predictors. **Results:** Among the nine algorithms, extreme gradient boosting (XGBoost) demonstrated the highest predictive performance. The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.884 (0.871–0.897) in the internal validation cohort and maintained comparable discrimination in the two external validation cohorts, eICU-Sepsis, derived from the eICU Collaborative Research Database (eICU-CRD), and DRECCv1.0 critical care database (DRECC)-Sepsis, with AUROCs of 0.877 (0.872–0.882) and 0.875 (0.856–0.894), respectively. Mechanical ventilation, comorbidity burden, lactate dehydrogenase, age, respiratory rate, and blood pressure made important contributions to the model's prediction of 28-day mortality. **Conclusion:** This interpretable machine learning model may facilitate early risk stratification of ICU patients with sepsis. Although the model shows potential for identifying high-risk individuals at the 24-hour prediction landmark, further prospective validation and clinical impact assessment are required before routine clinical implementation.

Keywords: sepsis; intensive care unit; 28-day mortality prediction; machine learning; risk stratification; explainable artificial intelligence

1. Introduction

Sepsis is a life-threatening syndrome caused by a dysregulated host response to infection and remains a major cause of death in critically ill patients. Despite advances in antimicrobial therapy, hemodynamic resuscitation, and organ support, short-term mortality remains high, particularly in severe cases [1,2,3]. Within the context of short-term outcomes, the 28-day mortality rate serves as a clinically significant endpoint that captures both early deaths due to shock and hypoperfusion, as well as later deaths attributable to persistent organ dysfunction, immune dysregulation, and secondary complications [4,5,6]. Early identification of patients at high risk of 28-day death is therefore essential for timely intervention and optimized allocation of critical care resources.

However, early prognostic assessment in sepsis is challenging due to the pronounced clinical heterogeneity of the syndrome [7,8]. Variations in infection sources, comorbidities, host responses, and trajectories of organ dysfunction often result in significant differences in outcomes, even among patients with similar initial clinical presentations [9]. Traditional scoring systems, such as the Sequential Organ Failure Assessment (SOFA), quick Sequential Organ Failure Assessment (qSOFA), and Acute Physiology and Chronic Health Evaluation II (APACHE II), are instrumental in assessing severity; however, they are constrained by a reliance on limited predefined variables and fixed scoring rules, which may inadequately capture the complex and nonlinear relationships inherent in sepsis prognosis [10,11,12].

Machine learning (ML) offers a potential solution to these limitations. By leveraging routinely available electronic health record data—including demographics, comorbidities, vital signs, laboratory results, and early treatment-related variables—ML models can more effectively characterize nonlinear associations and feature interactions compared to conventional scoring systems [13,14]. Recent research indicates that ML models may surpass traditional severity scores in predicting mortality associated with sepsis [15,16,17]. However, many published sepsis prediction models remain retrospective and single-center, while rigorous external or temporal validation across heterogeneous populations remains limited, thereby restricting their transportability and clinical applicability [17,18].

In this study, we aimed to develop and externally validate an interpretable machine learning model for predicting 28-day mortality in intensive care unit (ICU) patients with sepsis using three independent critical care cohorts. We also evaluated the model's performance and interpretability across the internal and external validation cohorts.

2. Methods

2.1 Study Population

This research utilized data from three databases. The MIMIC-Sepsis cohort, derived from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database [19], comprising 37,789 patients, served as the primary discovery cohort and was divided into an 8:2 ratio for model training and internal validation. For external validation, we utilized the eICU-Sepsis dataset [20], comprising 58,012 patients, which includes multicenter ICU patient records documenting clinical interventions and outcomes across the United States, as well as the DRECC-Sepsis dataset, consisting of 2432 patients, which represents a local hospital cohort of de-identified ICU patient records related to sepsis. The DRECC dataset was primarily developed based on data from Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, resulting in a comprehensive, multi-modal critical care database (DRECCv1.0). DRECC encompasses the complete patient information continuum from emergency admission to rehabilitation follow-up, incorporating multi-dimensional baseline data, temporally precise vital signs, comprehensive treatment records, various data modalities, and prognostic follow-up. For the purposes of this study, the three harmonized cohorts (MIMIC-Sepsis, eICU-Sepsis, and DRECC-Sepsis), comprising a total of 98,233 patients, were collectively referred to as the Multicenter Sepsis Database.

The inclusion criteria are as follows: ① adult ICU patients aged ≥18 years; ② patients meeting the Sepsis-3 criteria, defined as suspected infection with an acute increase in SOFA score of ≥2 points from 48 hours before to 24 hours after the onset of suspected infection. Suspected infection was identified using antimicrobial–culture pairing, defined as culture sampling within 24 hours after antimicrobial initiation or antimicrobial initiation within 72 hours after culture sampling [21,22]; and ③ an ICU stay of >24 hours. The exclusion criteria were as follows: ① missing critical outcome data, including 28-day mortality status (e.g., time of death) or ICU admission/discharge dates; and ② >10% missingness in key baseline variables, including vital signs (e.g., temperature, heart rate, respiratory rate, and mean arterial pressure) and laboratory parameters (e.g., lactate and renal/hepatic function markers). For each patient, the proportion of missing data was calculated as the number of missing candidate baseline variables divided by the total number of candidate baseline variables. Patients with a missing data proportion >10% were excluded.

The study workflow is shown in Fig. 1, encompassing data preprocessing, model development and interpretation, and external validation.

2.2 Study Outcome

The primary outcome was 28-day mortality after ICU admission. Patients who died within 28 days were classified as non-survivors, whereas the others were classified as survivors. ICU length of stay was calculated from ICU admission to ICU discharge or death, whichever occurred first. Only the first eligible ICU admission was included. Because the prediction landmark was set at 24 hours after ICU admission, model predictions were generated only for patients who remained in the ICU beyond this landmark; the 28-day outcome window was calculated from the time of ICU admission.

2.3 Feature Selection

A total of 45 variables were considered as candidate predictors based on a review of the relevant literature and their availability across the three databases [23,24,25,26,27,28]. A landmark prediction design was used, with the prediction landmark set at 24 hours after ICU admission. For vital signs and laboratory variables, the earliest valid value recorded from ICU admission to 24 hours after admission was extracted. Systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP) were extracted independently according to the earliest available valid record for each variable and were therefore not necessarily measured at the same timestamp. Baseline comorbidities were identified from diagnoses documented at or before ICU admission. Treatment-related variables included therapies initiated or administered during the same 24-hour predictor window and were therefore available at the prediction landmark. The 45 candidate predictors comprised two demographic variables, eight vital signs and bedside measurements, 26 laboratory variables, six comorbidity-related variables, and three early treatment-related variables. Demographic variables included age and sex. Vital signs and bedside measurements included systolic blood pressure, diastolic blood pressure, mean arterial pressure, heart rate, respiratory rate,

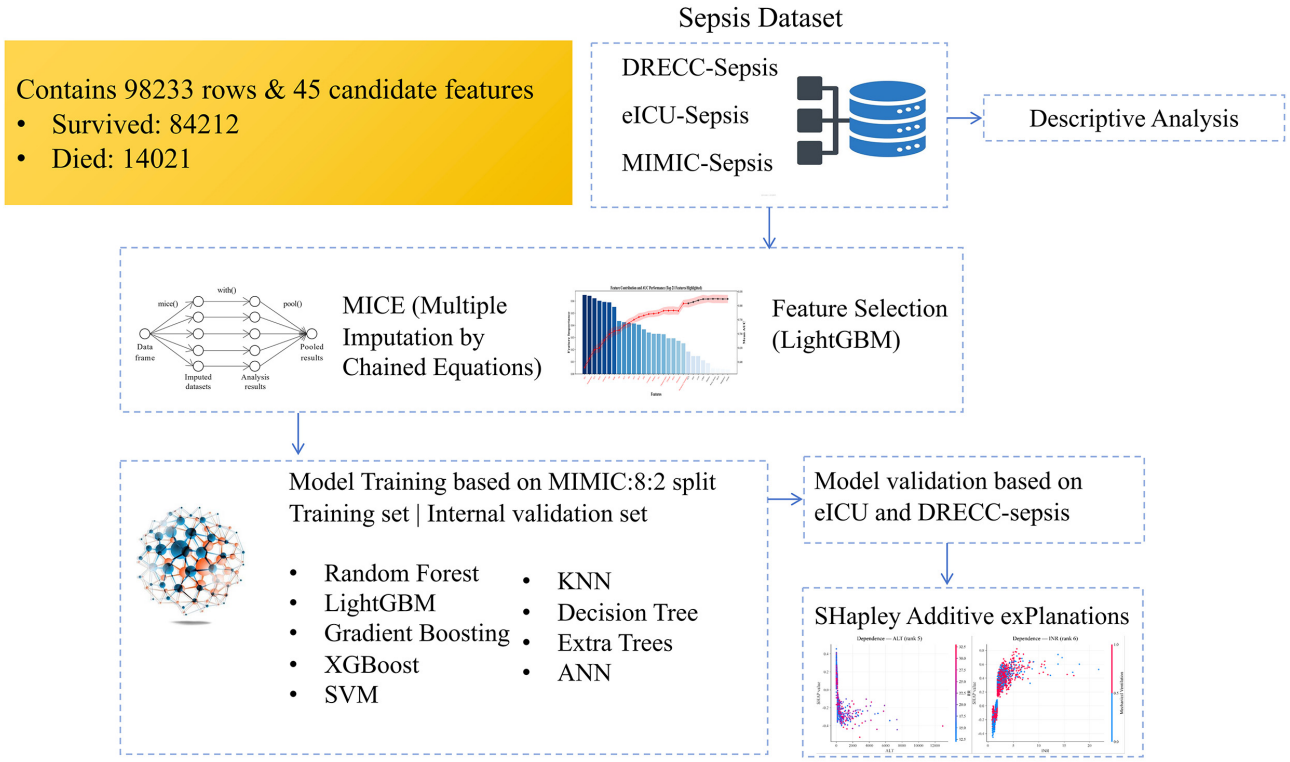

Fig. 1. Schematic overview of the study design. The diagram summarizes cohort construction, preprocessing, feature selection, model development, validation, and interpretation. Data from the MIMIC-Sepsis, eICU-Sepsis, and DRECC-Sepsis cohorts were screened to construct the sepsis dataset. Missing values were handled using multiple imputation by chained equations, candidate predictors were ranked and selected using LightGBM, and multiple machine learning models were trained using the MIMIC-Sepsis cohort. Model performance was evaluated in the internal validation cohort and in the external eICU-Sepsis and DRECC-Sepsis cohorts. Model interpretation was performed using SHAP-based analyses. Abbreviations: ANN, artificial neural network; DRECC, DRECCv1.0 critical care database; eICU, eICU Collaborative Research Database; LightGBM, light gradient boosting machine; MICE, multiple imputation by chained equations; MIMIC, Medical Information Mart for Intensive Care; SHAP, SHapley Additive exPlanations; SVM, support vector machine.

body temperature, peripheral oxygen saturation, and Glasgow Coma Scale score. Laboratory variables included glucose, platelet count, white blood cell count, hemoglobin, blood urea nitrogen, creatinine, sodium, potassium, chloride, bicarbonate, lactate, lactate dehydrogenase, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, gamma-glutamyl transferase, total bilirubin, direct bilirubin, C-reactive protein (CRP), international normalized ratio, activated partial thromboplastin time, D-dimer, pH, base excess, arterial partial pressure of oxygen (PaO_2), and the ratio of arterial oxygen partial pressure to the fraction of inspired oxygen (PaO_2/FiO_2 ratio). Comorbidity-related variables included the Charlson Comorbidity Index, myocardial infarction, congestive heart failure, cerebrovascular disease, chronic obstructive pulmonary disease, and renal disease. Early treatment-related variables included mechanical ventilation, continuous renal replacement therapy, and systemic corticosteroid exposure.

Comorbidities were identified from ICD-9/10 diagnosis codes documented at or before ICU admission. The Charlson Comorbidity Index was calculated using a coding algorithm by Quan et al. [29]. The Charlson Comorbidity Index component conditions shown in Table 1 were used to characterize baseline comorbidity burden. In addition to the Charlson Comorbidity Index total score, five components—myocardial infarction, congestive heart failure, cerebrovascular disease, chronic pulmonary disease, and renal disease—were included as independent candidate predictors; the remaining components were used for descriptive purposes only.

To minimize bias due to missingness while maximizing sample utilization, we excluded patients with >10% missing candidate baseline variables according to the prespecified quality-control rules. Remaining missing values were imputed using multiple imputation (MI). Multiple imputation was performed using the mice package in R (version 4.3.2; R Foundation for Statistical Computing, Vienna, Austria) after patient-level missingness exclusion and before feature selection. During cross-validation, imputation models were fitted using training-fold data only and were

Table 1. Baseline characteristics of participants in Multicenter Sepsis Database.

Characteristics		Total	eICU-Sepsis	MIMIC-Sepsis	DRECC-Sepsis	*p*-value	SMD
Number of patients		98,233	58,012	37,789	2432	N/A	N/A
Sex, *n* (%)	Female	44,339 (45.1)	27,624 (47.6)	15,881 (42.0)	834 (34.3)	<0.001	0.272
	Male	53,894 (54.9)	30,388 (52.4)	21,908 (58.0)	1598 (65.7)		
Age, mean ± SD, years		65.20 ± 16.43	64.44 ± 16.70	66.21 ± 15.99	67.52 ± 15.37	<0.001	0.185
Vital signs							
HR, mean ± SD, Bpm		92.96 ± 21.38	93.81 ± 21.64	91.40 ± 20.71	97.75 ± 23.56	<0.001	0.304
RR, mean ± SD, breaths/min		20.73 ± 6.78	21.16 ± 6.98	20.08 ± 6.42	21.29 ± 6.91	<0.001	0.188
SpO_2 (mean ± SD, %)		96.62 ± 4.69	96.51 ± 4.85	96.74 ± 4.47	96.74 ± 4.60	<0.001	0.049
MAP, mean ± SD, mmHg		79.34 ± 26.58	76.24 ± 29.43	83.81 ± 21.21	83.76 ± 18.19	<0.001	0.286
SBP, mean ± SD, mmHg		106.73 ± 30.13	122.91 ± 29.52	87.72 ± 16.89	122.07 ± 28.63	<0.001	1.927
DBP, mean ± SD, mmHg		57.57 ± 19.87	68.85 ± 19.15	44.54 ± 10.99	64.62 ± 15.94	<0.001	1.769
Temperature, mean ± SD, °C		37.12 ± 1.82	36.74 ± 2.41	37.48 ± 0.81	37.27 ± 1.00	<0.001	0.410
GCS, mean ± SD		13.24 ± 3.14	12.57 ± 3.50	14.11 ± 2.29	11.99 ± 4.01	<0.001	0.876
Laboratory tests							
Hb, mean ± SD, g/dL		10.87 ± 2.48	11.31 ± 2.53	10.31 ± 2.25	9.28 ± 2.55	<0.001	0.804
PLT, median (IQR), $\times 10^9$/L		199.00 [138.00–274.00]	211.00 [151.00–285.00]	185.00 [126.00–259.00]	129.00 [73.00–205.00]	<0.001	0.686
WBC, median (IQR), $\times 10^9$/L		11.80 [8.04–16.70]	12.20 [8.40–17.30]	11.30 [7.80–15.90]	9.10 [5.80–14.60]	<0.001	0.316
BUN, median (IQR), mg/dL		23.00 [15.00–39.00]	23.00 [14.00–39.00]	22.00 [15.00–38.00]	23.80 [14.28–41.38]	0.273	0.049
Chloride, mean ± SD, mmol/L		100.36 ± 14.94	97.96 ± 18.08	103.70 ± 7.24	106.10 ± 6.34	<0.001	0.458
Sodium, mean ± SD, mmol/L		137.74 ± 5.97	137.30 ± 6.09	138.27 ± 5.69	139.91 ± 6.04	<0.001	0.430
Potassium, mean ± SD, mmol/L		4.23 ± 0.81	4.21 ± 0.82	4.26 ± 0.79	4.16 ± 0.66	<0.001	0.127
Lactate, median (IQR), mmol/L		2.00 [1.20–3.30]	2.00 [1.20–3.40]	2.00 [1.40–3.20]	1.50 [1.00–2.30]	<0.001	0.264
AST, median (IQR), U/L		35.00 [21.00–72.00]	31.00 [20.00–60.00]	44.00 [25.00–102.00]	35.00 [23.00–65.75]	<0.001	0.146
ALT, median (IQR), U/L		27.00 [16.00–52.00]	26.00 [17.00–47.00]	29.00 [17.00–67.00]	25.00 [14.00–49.00]	<0.001	0.156
Creatinine, median (IQR), mg/dL		1.20 [0.80–2.00]	1.15 [0.80–1.94]	1.20 [0.80–2.00]	0.94 [0.59–1.76]	<0.001	0.173
Glucose, median (IQR), mg/dL		137.00 [110.00–184.00]	131.00 [105.00–175.00]	146.00 [118.00–198.00]	136.80 [111.24–187.20]	<0.001	0.184
INR, median (IQR)		1.30 [1.10–1.69]	1.20 [1.10–1.60]	1.40 [1.20–1.70]	1.23 [1.11–1.41]	<0.001	0.311
Bicarbonate, mean ± SD, mmol/L		24.36 ± 5.89	24.50 ± 5.86	21.79 ± 6.28	23.58 ± 5.76	<0.001	0.461
APTT, median (IQR), s		33.20 [28.60–43.00]	31.00 [27.20–36.20]	33.60 [28.80–45.40]	43.80 [37.90–54.00]	<0.001	1.005
Total bilirubin, median (IQR), mg/dL		0.70 [0.40–1.30]	0.60 [0.40–1.00]	0.70 [0.40–1.70]	0.88 [0.61–1.51]	<0.001	0.311
pH, mean ± SD		7.33 ± 0.12	7.36 ± 0.11	7.30 ± 0.11	7.39 ± 0.09	<0.001	0.815
CRP, median (IQR), mg/L		43.90 [8.70–123.80]	59.40 [17.77–164.03]	30.60 [5.30–99.50]	81.80 [24.80–145.00]	<0.001	0.484
P/F ratio, median (IQR)		182.00 [103.95–276.19]	216.00 [142.65–305.00]	150.00 [88.00–246.00]	266.00 [198.00–347.00]	<0.001	0.834
LDH, median (IQR), IU/L		296.00 [211.00–473.00]	280.00 [199.00–468.00]	303.00 [218.00–487.00]	286.00 [206.00–393.00]	<0.001	0.155
ALP, median (IQR), U/L		92.00 [67.00–134.00]	91.00 [67.00–129.00]	93.00 [66.00–144.00]	96.00 [65.00–144.00]	<0.001	0.159
GGT, median (IQR), U/L		56.00 [29.00–119.00]	N/A	96.00 [39.00–246.00]	54.00 [29.00–109.00]	<0.001	0.745

Table 1. Continued.

Characteristics		Total	eICU-Sepsis	MIMIC-Sepsis	DRECC-Sepsis	*p*-value	SMD
Base excess, median (IQR), mmol/L		0.00 [−3.00–3.00]	0.00 [−4.90–3.00]	0.00 [−2.00–3.00]	–1.10 [−4.10–2.00]	<0.001	0.335
PaO_2, median (IQR), mmHg		80.00 [56.00–119.00]	88.00 [67.00–143.00]	65.00 [40.00–97.00]	107.00 [87.00–135.00]	<0.001	0.658
D-dimer, median (IQR), µg/mL		2.24 [1.17–4.90]	N/A	2.07 [1.02–5.27]	2.33 [1.29–4.81]	0.167	0.032
Direct bilirubin, median (IQR), mg/dL		0.20 [0.10–0.50]	0.20 [0.10–0.50]	0.12 [0.05–0.26]	0.29 [0.17–0.65]	<0.001	1.077
Comorbidities							
Myocardial infarction, *n* (%)	No	89,123 (90.7)	55,783 (96.2)	30,927 (81.8)	2413 (99.2)	<0.001	0.703
	Yes	9110 (9.3)	2229 (3.8)	6862 (18.2)	19 (0.8)		
Congestive heart failure, *n* (%)	No	77,974 (79.4)	50,563 (87.2)	25,165 (66.6)	2246 (92.4)	<0.001	0.672
	Yes	20,259 (20.6)	7449 (12.8)	12,624 (33.4)	186 (7.6)		
Peripheral vascular disease, *n* (%)	No	92,510 (94.2)	57,309 (98.8)	33,051 (87.5)	2150 (88.4)	<0.001	0.503
	Yes	5723 (5.8)	703 (1.2)	4738 (12.5)	282 (11.6)		
Cerebrovascular disease, *n* (%)	No	89,914 (91.5)	55,575 (95.8)	32,228 (85.3)	2111 (86.8)	<0.001	0.375
	Yes	8319 (8.5)	2437 (4.2)	5561 (14.7)	321 (13.2)		
Dementia, *n* (%)	No	94,142 (95.8)	55,846 (96.3)	35,885 (95.0)	2411 (99.1)	<0.001	0.267
	Yes	4091 (4.2)	2166 (3.7)	1904 (5.0)	21 (0.9)		
COPD, *n* (%)	No	76,796 (78.2)	47,130 (81.2)	27,370 (72.4)	2296 (94.4)	<0.001	0.628
	Yes	21,437 (21.8)	10,882 (18.8)	10,419 (27.6)	136 (5.6)		
Rheumatic disease, *n* (%)	No	96,420 (98.2)	57,691 (99.4)	36,374 (96.3)	2355 (96.8)	<0.001	0.241
	Yes	1813 (1.8)	321 (0.6)	1415 (3.7)	77 (3.2)		
Peptic ulcer disease, *n* (%)	No	96,319 (98.1)	57,515 (99.1)	36,472 (96.5)	2332 (95.9)	<0.001	0.223
	Yes	1914 (1.9)	497 (0.9)	1317 (3.5)	100 (4.1)		
Mild liver disease, *n* (%)	No	88,873 (90.5)	55,575 (95.8)	31,661 (83.8)	1637 (67.3)	<0.001	0.804
	Yes	9360 (9.5)	2437 (4.2)	6128 (16.2)	795 (32.7)		
Diabetes, *n* (%)	No	83,056 (84.5%)	55,856 (96.3%)	25,282 (66.9%)	1918 (78.9%)	<0.001	0.664
	Yes	15,177 (15.5%)	2156 (3.7%)	12,507 (33.1%)	514 (21.1%)		
Spinal cord injury, *n* (%)	No	95,909 (97.6%)	57,689 (99.4%)	35,795 (94.7%)	2425 (99.7)	<0.001	0.356
	Yes	2324 (2.4%)	323 (0.6%)	1994 (5.3%)	7 (0.3)		
Renal disease, *n* (%)	No	79,425 (80.9%)	50,679 (87.4%)	27,660 (73.2%)	1086 (44.7%)	<0.001	0.951
	Yes	18,808 (19.1%)	7333 (12.6%)	10,129 (26.8%)	1346 (55.3%)		
Malignancy, *n* (%)	No	88,367 (90.0%)	54,063 (93.2%)	32,375 (85.7%)	1929 (79.3%)	<0.001	0.416
	Yes	9866 (10.0%)	3949 (6.8%)	5414 (14.3%)	503 (20.7%)		
Severe liver disease, *n* (%)	No	93,645 (95.3%)	57,018 (98.3%)	34,547 (91.4%)	2080 (85.5%)	<0.001	0.518
	Yes	4588 (4.7%)	994 (1.7%)	3242 (8.6%)	352 (14.5%)		
Metastatic solid tumor, *n* (%)	No	95,080 (96.8%)	57,279 (98.7%)	35,383 (93.6%)	2418 (99.4%)	<0.001	0.358
	Yes	3153 (3.2%)	733 (1.3%)	2406 (6.4%)	14 (0.6%)		
AIDS, *n* (%)	No	97,762 (99.5%)	57,879 (99.8%)	37,460 (99.1%)	2423 (99.6%)	<0.001	0.091
	Yes	471 (0.5%)	133 (0.2%)	329 (0.9%)	9 (0.4%)		

Table 1. Continued.

Characteristics		Total	eICU-Sepsis	MIMIC-Sepsis	DRECC-Sepsis	*p*-value	SMD
Charlson Comorbidity Index, median (IQR)		4 [3–6]	4 [3–5]	5 [3–7]	3 [2–5]	<0.001	0.702
Therapy							
CRRT, *n* (%)	No	94,293 (96.0)	57,066 (98.4)	35,037 (92.7)	2190 (90.0)	<0.001	0.386
	Yes	3940 (4.0)	946 (1.6)	2752 (7.3)	242 (10.0)		
Mechanical ventilation, *n* (%)	No	54,329 (55.3)	36,279 (62.5)	16,950 (44.9)	1100 (45.2)	<0.001	0.357
	Yes	43,904 (44.7)	21,733 (37.5)	20,839 (55.1)	1332 (54.8)		
Glucocorticoids, *n* (%)	No	95,021 (96.7)	57,839 (99.7)	35,634 (94.3)	1548 (63.7)	<0.001	1.185
	Yes	3212 (3.3)	173 (0.3)	2155 (5.7)	884 (36.3)		
Prognosis							
28-day Mortality, *n* (%)	Alive	84,212 (85.7)	50,477 (87.0)	31,917 (84.5)	1818 (74.8)	<0.001	0.315
	Dead	14,021 (14.3)	7535 (13.0)	5872 (15.5)	614 (25.2)		
ICU length of stay, median (IQR), days		2.60 [1.33–5.27]	2.23 [1.16–4.49]	3.00 [1.58–6.34]	7.43 [3.15–16.39]	<0.001	1.383

Notes: All comorbidities listed in Table 1 were used to characterize baseline comorbidity burden. In addition to the Charlson Comorbidity Index total score, myocardial infarction, congestive heart failure, cerebrovascular disease, chronic pulmonary disease, and renal disease were included as independent candidate predictors; the remaining Charlson Comorbidity Index components were used for descriptive purposes only. GGT and D-dimer were unavailable in the eICU-Sepsis cohort; their *p*-values therefore represent comparisons between the MIMIC-Sepsis and DRECC-Sepsis cohorts. Table 1 presents pre-imputation data. SBP, DBP, and MAP were extracted independently as the earliest valid measurements within the first 24 hours after ICU admission and were not necessarily derived from the same blood-pressure measurement episode; therefore, cohort-level MAP values may differ from values approximated using the reported SBP and DBP summary statistics.

Abbreviations: AIDS, acquired immunodeficiency syndrome; ALP, alkaline phosphatase; ALT, alanine aminotransferase; APTT, activated partial thromboplastin time; AST, aspartate aminotransferase; BUN, blood urea nitrogen; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; CRRT, continuous renal replacement therapy; DBP, diastolic blood pressure; GCS, Glasgow Coma Scale; GGT, gamma-glutamyl transferase; Hb, hemoglobin; HR, heart rate; ICU, intensive care unit; INR, international normalized ratio; LDH, lactate dehydrogenase; MAP, mean arterial pressure; P/F ratio, PaO_2/FiO_2 ratio; PaO_2, arterial partial pressure of oxygen; PLT, platelet count; RR, respiratory rate; SBP, systolic blood pressure; SD, standard deviation; SpO_2, peripheral oxygen saturation; WBC, white blood cell count.

subsequently applied to the corresponding validation fold to minimize information leakage. Variable-specific missingness rates and full details of the imputation procedure are provided in **Supplementary Table 1**. To rank the candidate predictors and select a parsimonious feature subset, we applied the LightGBM algorithm exclusively within the training cohort. Feature importance was quantified in terms of split-based importance, defined as the number of times each predictor was used to split a node across all trees. Candidate predictors were ranked in descending order of importance. Starting with the highest-ranked predictor, variables were added sequentially, and the mean area under the receiver operating characteristic curve (AUROC) was evaluated using 5-fold cross-validation in the training cohort. The optimal number of predictors was defined as the smallest feature subset whose mean five-fold cross-validated AUROC was within 0.005 of the maximum mean AUROC observed across all evaluated feature subsets. Based on this criterion, 16 predictors were retained for model development. All screening steps were performed exclusively on the training set with fixed parameters. The same parameters were applied in the validation set to prevent information leakage. The "scikit-learn" library in Python (version 3.11; Python Software Foundation, Beaverton, OR, USA) was utilized for this research.

2.4 Model Development and Comparison

Using the 16 predictors retained after LightGBM-based feature selection, the MIMIC-Sepsis cohort was used as the development cohort and randomly divided into a training set and an internal validation set at an 8:2 ratio using stratified sampling according to 28-day mortality status. All preprocessing, imputation, feature selection, and hyperparameter tuning procedures were performed exclusively within the training set. The fitted preprocessing parameters and final model specifications were subsequently applied without refitting to the internal validation set and the two external validation cohorts, eICU-Sepsis and DRECC-Sepsis, to minimize information leakage. Five-fold cross-validation was uniformly employed for model training within the training set: in each fold, 4 folds were used to fit the model, while the remaining fold was used to evaluate model performance, with the mean AUROC serving as the optimization objective for hyperparameter selection. Details of the hyperparameter optimization procedure, including the search method, parameter search ranges, and final selected hyperparameters for each candidate algorithm, are provided in **Supplementary Table 2**. Given the imbalance between survivors and non-survivors, no random oversampling, undersampling, or synthetic sample generation was applied. Class-weighting options, where supported and prespecified, were evaluated exclusively within the training-fold hyperparameter search. The internal and external validation cohorts retained their observed event rates and were not resampled. The primary comparison between models was based on AUROC, which does not require selection of a single classification threshold. For threshold-dependent performance measures, the classification cutoff was determined using the Youden J statistic based exclusively on out-of-fold predicted probabilities from the training set. Once selected, this cutoff was locked and applied without re-optimization to the internal validation set and both external validation cohorts. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and Kappa were subsequently calculated using this fixed cutoff.

Based on this framework, nine mainstream models were constructed: random forest, LightGBM, gradient boosting, extreme gradient boosting (XGBoost), support vector machine (SVM), k-nearest neighbors (KNN), decision tree, extra trees, and artificial neural network (ANN). These models were then evaluated for generalization on two independent external validation datasets: eICU-Sepsis and DRECC-Sepsis. Beyond AUROC, we also evaluated other threshold-related metrics (see above).

2.5 Model Interpretation

To enhance model transparency, this study employed the SHAP (SHapley Additive exPlanations) method to improve model interpretability [30,31]. SHAP aims to explain the "black box" decision-making process of complex models by calculating each feature's contribution to a single prediction, expressed as a SHAP value. It provides both global and local explanations: globally quantifying the overall contribution of each feature, and locally elucidating the specific reasons behind predictions for individual subjects. Global feature importance was quantified using mean absolute SHAP values. Beeswarm plots were used to show the direction and magnitude of feature contributions across patients, with color indicating the corresponding feature value. SHAP dependence plots were generated for selected predictors to examine potential nonlinear relationships between feature values and model output. This study was implemented using the scikit-learn and SHAP libraries in Python 3.11.

2.6 Statistical Analysis

Baseline characteristics across the three cohorts were summarized using R (version 4.3.2; R Foundation for Statistical Computing, Vienna, Austria) and Python (version 3.11; Python Software Foundation, Beaverton, OR, USA) Normality of continuous variables was assessed using the Kolmogorov–Smirnov test. Homogeneity of variances for normally distributed continuous variables was assessed using Levene's test. Normally distributed continuous variables were presented as mean ± standard deviation (SD), whereas non-normally distributed continuous variables were expressed as median (interquartile range [IQR]). Categorical variables were summarized as counts and proportions (n [%]). For comparisons across the three

cohorts, normally distributed continuous variables were analyzed using one-way analysis of variance (ANOVA) when the homogeneity-of-variance assumption was satisfied and Welch's ANOVA when this assumption was violated. Non-normally distributed continuous variables were compared using the Kruskal–Wallis test, and categorical variables were analyzed using the chi-square test. For variables available in only two cohorts (gamma-glutamyl transferase (GGT) and D-dimer), the Mann–Whitney *U* test was used because these variables were non-normally distributed. Standardized mean differences (SMDs) were also calculated to quantify inter-cohort heterogeneity, with an absolute SMD <0.10 considered indicative of a negligible difference. A two-sided p-value < 0.05 was considered statistically significant.

3. Results

3.1 Participant Characteristics

The Multicenter Sepsis Database comprised 98,233 patients diagnosed with sepsis, with a mean age of 65.20 ± 16.43 years. Of these patients, 53,894 (54.9%) were male and 44,339 (45.1%) were female, resulting in a male-to-female ratio of 1.215:1. Table 1 provides descriptive statistics for key baseline characteristics and outcome measures across the multicenter datasets. Notably, apart from D-dimer and BUN, the three cohorts exhibited variations in demographic characteristics, comorbidity burdens, physiological and laboratory parameters measured during the first 24 hours after ICU admission, treatment exposures, and clinical outcomes, underscoring the substantial heterogeneity present in the real-world population.

3.2 Feature Extraction and Selection

This study utilized the LightGBM model to evaluate and rank the importance of all candidate features. To evaluate the impact of the number of features on performance and to determine an optimal subset, recursive feature addition was employed. As illustrated in Fig. 2, the mean AUROC increased rapidly as the highest-ranked predictors were added and improved more gradually thereafter. According to the selection criterion, the smallest feature subset whose mean five-fold cross-validated AUROC was within 0.005 of the maximum mean AUROC observed across all evaluated feature subsets contained 16 predictors. These predictors were mechanical ventilation, Charlson Comorbidity Index, lactate dehydrogenase, age, respiratory rate, systolic blood pressure, activated partial thromboplastin time, body temperature, white blood cell count, creatinine, platelet count, PaO_2, heart rate, glucose, PaO_2/FiO_2 ratio, and mean arterial pressure.

3.3 Model Development and Performance Comparison

Using the 16 features, we trained nine machine-learning models. The findings indicate that XGBoost exhibited superior performance in both the training and internal validation datasets (**Supplementary Tables 3,4**). Specifically, within the training dataset, XGBoost achieved an accuracy of 88.86%, sensitivity of 71.82%, specificity of 92.00%, positive predictive value (PPV) of 62.29%, negative predictive value (NPV) of 94.66%, F1 score of 66.71%, and a Kappa value of 0.601. In the internal validation dataset, the corresponding metrics were 88.07%, 70.44%, 91.31%, 59.84%, 94.38%, 64.71%, and 0.576, respectively. The consistency of metrics between the training and validation datasets suggests robust stability with minimal overfitting. In terms of AUROC, XGBoost also demonstrated the best performance (Fig. 3A,B). Specifically, it achieved an AUROC of 0.890 (0.884–0.896) on the training dataset and 0.884 (0.871–0.897) on the internal validation dataset, with LightGBM closely following.

3.4 External Validation

External validation was performed in the eICU-Sepsis and DRECC-Sepsis cohorts to assess model generalizability. XGBoost again performed best in both external datasets (**Supplementary Tables 5,6**). In eICU-Sepsis, accuracy was 87.81%, sensitivity 69.12%, specificity 90.60%, PPV 52.33%, NPV 95.16%, F1 score 59.56%, and Kappa value 0.525. In DRECC-Sepsis, the corresponding metrics were 84.54%, 68.73%, 89.88%, 69.64%, 89.49%, 69.18%, and 0.589. For AUROC, XGBoost achieved 0.877 (0.872–0.882) and 0.875 (0.856–0.894) in the eICU-Sepsis and DRECC-Sepsis cohorts, respectively, whereas LightGBM achieved 0.863 and 0.861 (Fig. 3C,D). These findings indicate that the model maintained comparable discrimination across the two external validation cohorts.

3.5 Model Explanation

To enhance the clinical interpretability of the model, we employed SHAP to analyze the final XGBoost model (Fig. 4). SHAP analysis indicated that the features with the greatest contributions to model prediction largely reflected illness severity and organ dysfunction. Mechanical ventilation showed the greatest contribution to model prediction, followed by the Charlson Comorbidity Index, lactate dehydrogenase (LDH), age, respiratory rate (RR), systolic blood pressure (SBP), activated partial thromboplastin time (APTT), body temperature, and white blood cell count (WBC). Higher values of Charlson Comorbidity Index, age, LDH, RR, WBC, heart rate, creatinine, APTT, and glucose, as well as the use of mechanical ventilation, generally contributed to higher predicted 28-day mortality risk. Lower values of platelet count (PLT), SBP, and the PaO_2/FiO_2 ratio generally contributed to higher predicted 28-day mortality risk. In contrast, the contribution of PaO_2 varied across patients and did not show a clear monotonic direction; therefore, PaO_2 was not interpreted as having a uniformly protective or harmful association with mortality risk. Hypothermia showed a stronger positive contribution to predicted mortality than hyperthermia.

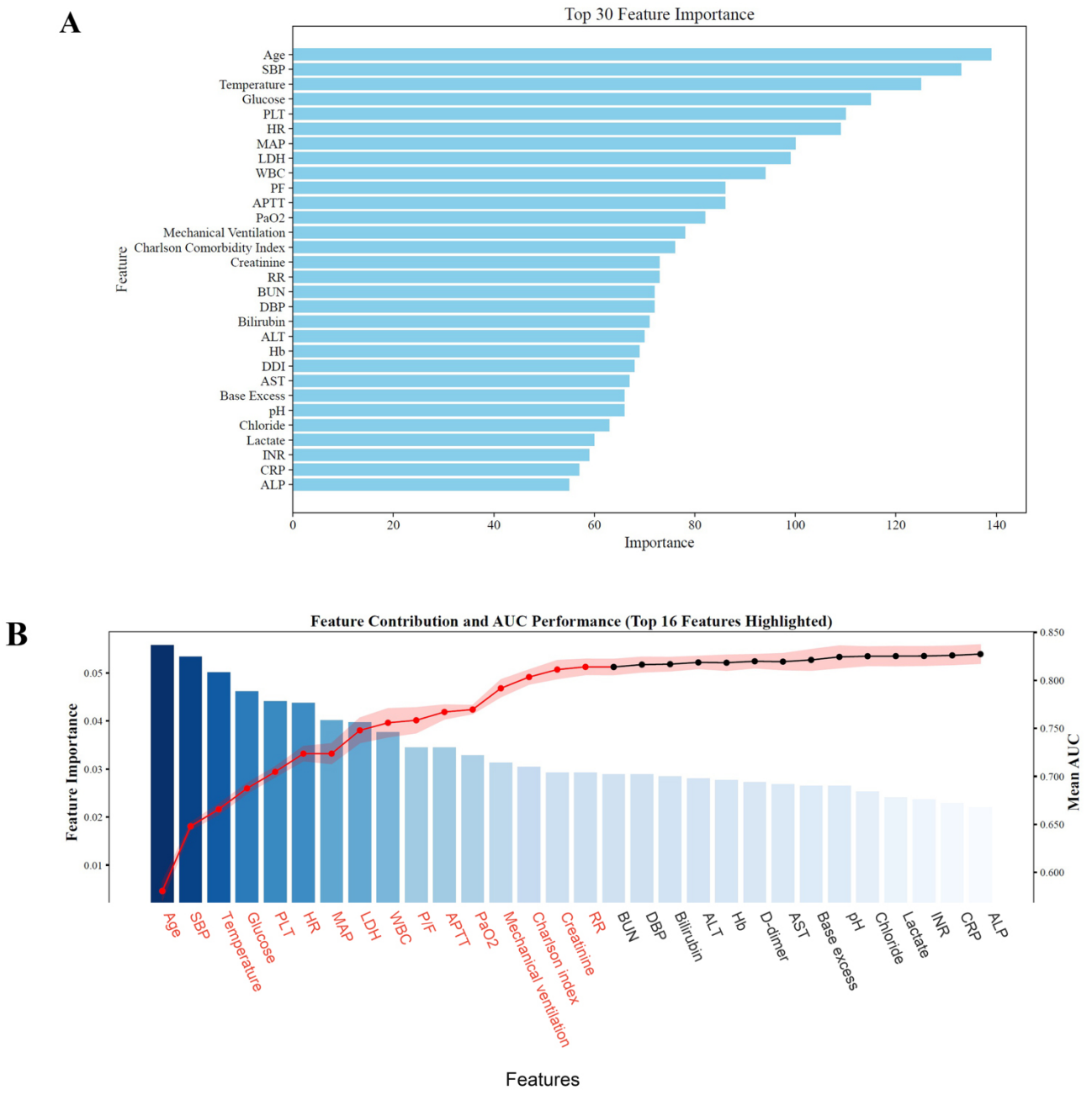

Fig. 2. Feature importance and stepwise feature contribution during feature selection. (A) Ranking of the top 30 candidate features according to importance estimated by the LightGBM model. Higher values indicate greater contribution to model discrimination. (B) Stepwise evaluation of feature contribution and mean AUROC as features were added sequentially according to their importance ranking. The first 16 features, highlighted in the plot, were retained because they constituted the smallest subset whose mean five-fold cross-validated AUROC was within 0.005 of the maximum mean AUROC observed across all evaluated feature subsets. The bars represent normalized relative feature importance, calculated from the raw split-based importance values, whereas the line represents the mean five-fold cross-validated AUROC obtained for each incrementally expanded feature subset. Abbreviations: AUROC, area under the receiver operating characteristic curve; LightGBM, light gradient boosting machine.

To further examine the nonlinear associations between key predictors and model output, SHAP dependence plots were generated (Fig. 5). Mechanical ventilation was associated with higher SHAP values, indicating a greater contribution to predicted mortality risk. Higher Charlson Comorbidity Index, age, and respiratory rate generally corresponded to increasing contributions to predicted risk. LDH showed a rapid increase in SHAP contribution at lower-to-moderate elevations, followed by a relative plateau at higher values. For SBP, very low values were associated

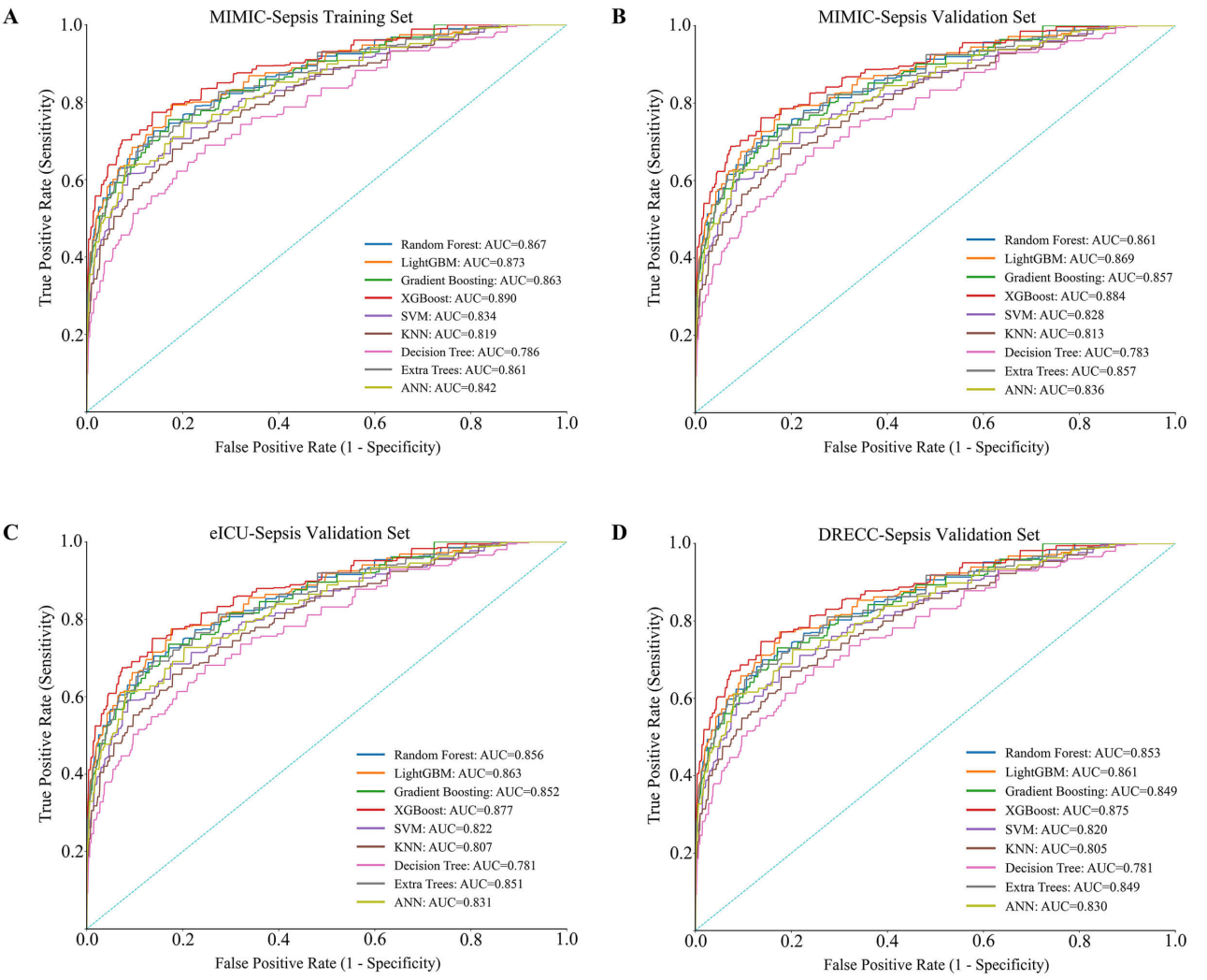

Fig. 3. Discriminative performance of machine learning models for 28-day mortality prediction in patients with sepsis. Receiver operating characteristic (ROC) curves compare nine candidate models across the model-development and validation cohorts. (A) MIMIC-Sepsis training set. (B) MIMIC-Sepsis internal validation set. (C) eICU-Sepsis external validation cohort. (D) DRECC-Sepsis external validation cohort. Across the four cohorts, XGBoost showed the highest or near-highest AUC, supporting its selection as the final prediction model. Abbreviations: ANN, artificial neural network; AUROC, area under the ROC curve; DRECC, DRECCv1.0 critical care database; eICU, eICU Collaborative Research Database; KNN, k-nearest neighbors; MIMIC, Medical Information Mart for Intensive Care; ROC, receiver operating characteristic; SVM, support vector machine; XGBoost, extreme gradient boosting.

with markedly increased predicted risk, whereas the SHAP contribution decreased as SBP approached an intermediate range and became more variable at higher values.

4. Discussion

This study developed and externally validated an interpretable XGBoost model for predicting 28-day mortality in ICU patients with sepsis using three independent critical care cohorts. The model achieved an AUROC of 0.884 in the internal validation cohort and maintained similar discrimination in the eICU-Sepsis and DRECC-Sepsis external validation cohorts, with AUROCs of 0.877 and 0.875, respectively (**Supplementary Table 7**). These findings suggest potential utility for early risk stratification, although prospective validation and clinical impact assessment are required before routine implementation.

Recent systematic reviews and meta-analyses have established machine learning algorithms as some of the most frequently utilized and best-performing methodologies. For example, Hou et al. [32] assessed the predictive performance of XGBoost for 30-day mortality risk in sepsis patients using the MIMIC-III retrospective cohort (n = 4559). The model achieved an AUROC of 0.857, surpassing the performance of both the SAPS II and logistic regression models. The decision curve and clinical impact curve analyses further suggested its potential clinical applicabil-

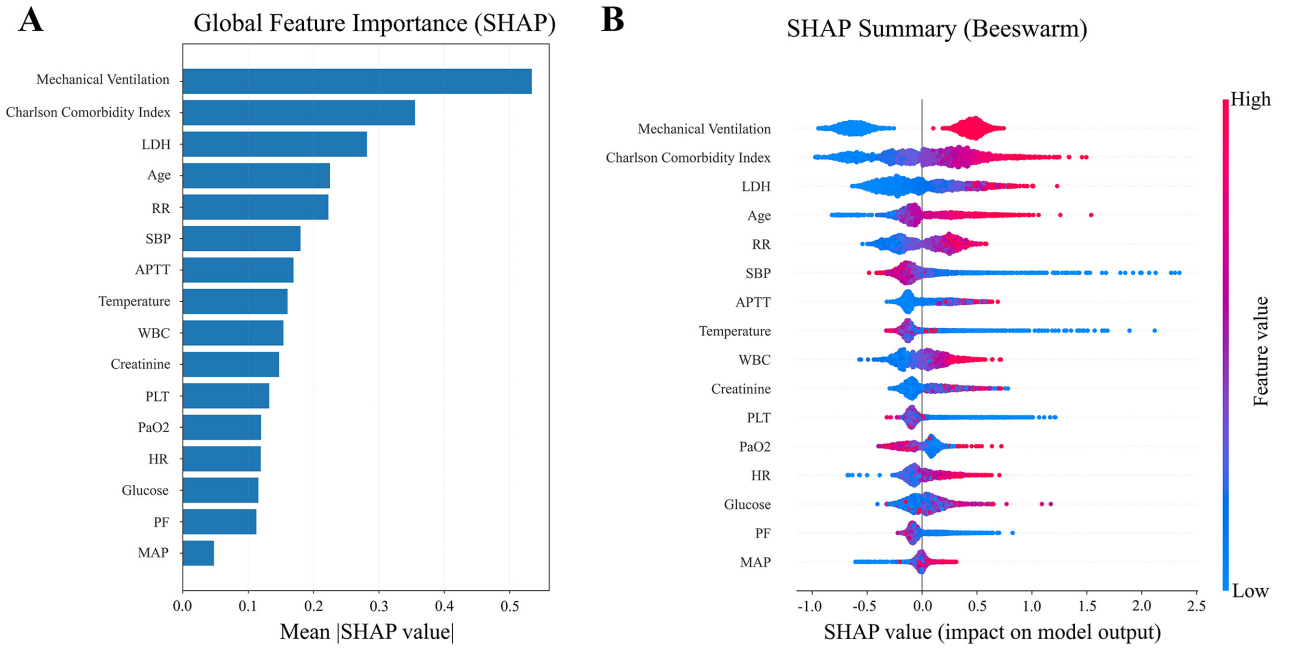

Fig. 4. Global interpretation of the final XGBoost model using SHAP. (A) Ranking of the top 16 predictors according to mean absolute SHAP values. (B) SHAP summary plot showing the direction and magnitude of feature contributions across patients. Abbreviations: APTT, activated partial thromboplastin time; LDH, lactate dehydrogenase; MAP, mean arterial pressure; PaO_2, arterial partial pressure of oxygen; P/F, PaO_2/FiO_2 ratio; PLT, platelet count; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; WBC, white blood cell count; XGBoost, extreme gradient boosting.

ity [32]. Similarly, in the MIMIC-IV dataset, Gao et al. [23] implemented a random forest model to predict short-term mortality in sepsis patients, achieving an AUROC of 0.94, which exceeded the performance of traditional scores such as SOFA. Although these models generally demonstrate superior discriminative ability compared with traditional scoring systems, they predominantly originate from single-center studies and are largely limited to internal validation. Therefore, their generalizability across diverse populations and clinical settings, as well as their practical application value, necessitates validation through prospective, multicenter studies [17,33,34].

External validation is essential for evaluating a model's generalization capacity and clinical applicability. Only through validation using independent datasets from different regions, populations, and healthcare systems can a model demonstrate its stability and reliability in real-world settings [17,35,36]. Based on a large cohort of 98,233 patients, this study evaluated 45 candidate variables available at the 24-hour prediction landmark and compared nine machine-learning algorithms. These variables are routinely collected in critical care settings and are generally available in electronic health records, which facilitates their use for cross-center model development and validation. However, differences in measurement practices, data capture, and patient characteristics across institutions may result in substantial variability in their distributions and should be considered when interpreting model transportability. Building upon this foundation, we sequentially constructed and compared nine mainstream machine learning models to identify the optimal algorithm. Results demonstrated that XGBoost achieved the best overall performance in both the eICU-Sepsis and DRECC-Sepsis independent external validation cohorts. These results indicate that the model possesses a certain degree of cross-center transferability and has potential to support early risk stratification of sepsis patients admitted to the ICU.

Despite the comparable discriminative performance of the model observed across the external validation cohorts, there was substantial inter-database heterogeneity, particularly in Charlson Comorbidity Index, ICU length of stay, glucocorticoid exposure, mechanical ventilation, and several comorbidities (see Table 1). These differences likely reflect variations in case mix, clinical practice, data capture, and healthcare systems across the MIMIC, eICU, and DRECC cohorts. Although predictor definitions and measurement units were harmonized before analysis, residual heterogeneity may remain because data collection and measurement practices differ across databases. Therefore, the preservation of AUROC across the external cohorts supports the transportability of model discrimination, but does not imply complete equivalence of model performance across settings. Differences in baseline risk and predictor distributions may still affect calibration and clinical performance, and further prospective validation in additional populations is warranted.

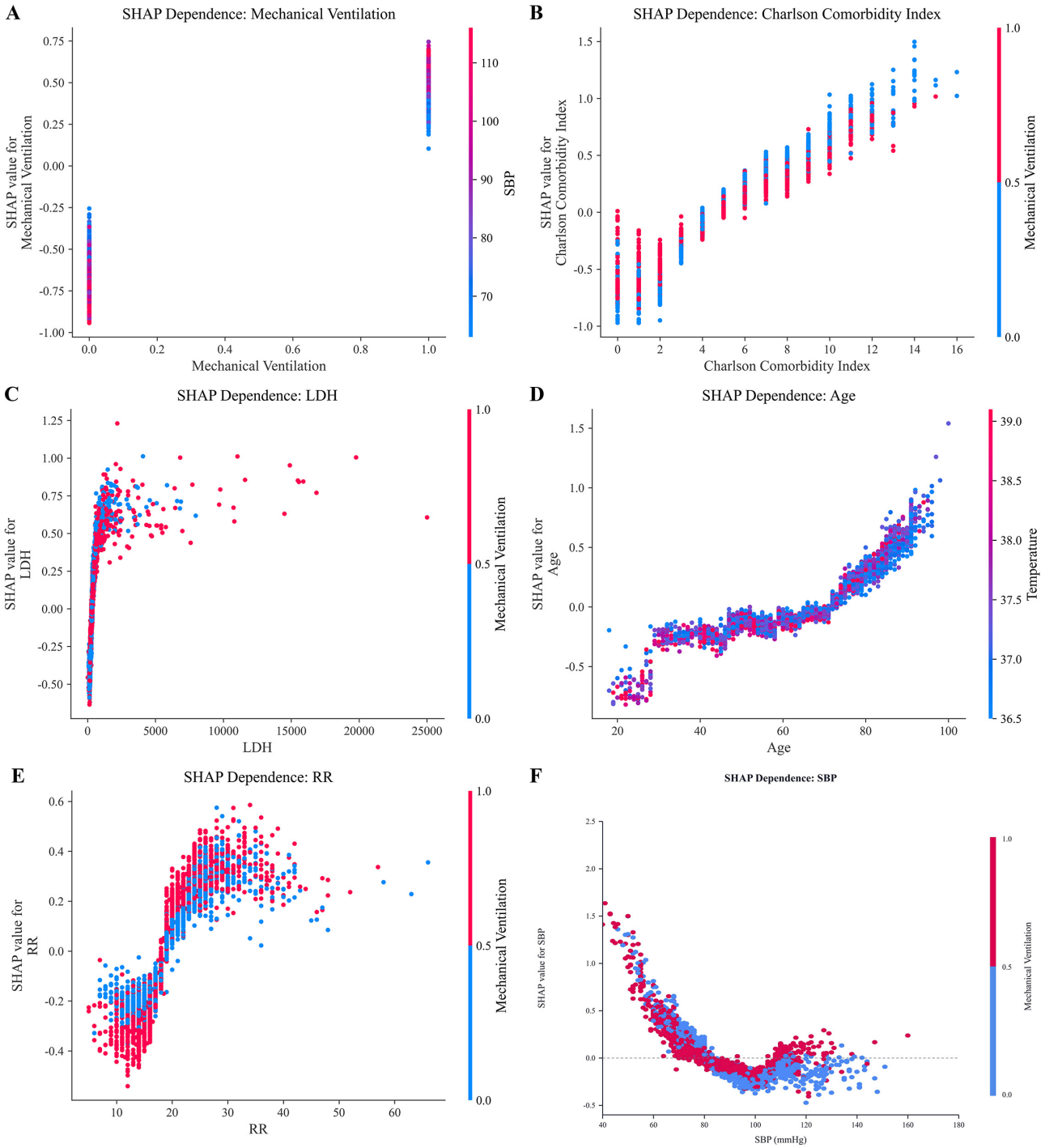

Fig. 5. SHAP dependence plots for selected predictors in the final XGBoost model. (A) Mechanical ventilation. (B) Charlson Comorbidity Index. (C) Lactate dehydrogenase. (D) Age. (E) Respiratory rate. (F) Systolic blood pressure. The x-axis represents the observed value of each predictor, and the y-axis represents its corresponding SHAP value. Points are colored according to the value of the interacting feature indicated by the color bar in each panel. Abbreviations: LDH, lactate dehydrogenase; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; XGBoost, extreme gradient boosting.

After conducting external validation, we employed SHAP to interpret the optimal XGBoost model at both global and individual levels. The global SHAP ranking identified mechanical ventilation, Charlson Comorbidity Index, lactate dehydrogenase (LDH), age, respiratory rate (RR), systolic blood pressure (SBP), and activated partial thromboplastin time (APTT) as important contributors to the model's prediction of 28-day mortality risk. These findings are biologically plausible and align with existing evidence: mechanical ventilation indicates respiratory fail-

ure and disease severity, correlating with increased mortality [37,38]; a higher Charlson comorbidity burden and advanced age suggest diminished physiological reserve and elevated baseline risk [39,40,41,42,43]; LDH serves as a marker of tissue injury and systemic inflammatory activation, which are associated with adverse outcomes in critical illness [44,45]. An elevated respiratory rate reflects acute physiological stress and is one of the three components of the qSOFA score, together with SBP and altered mentation [46]. Recent SHAP-based research similarly highlighted comorbidity burden, acute physiological dysregulation (e.g., hypothermia, tachypnea), and laboratory abnormalities (e.g., coagulopathy) as important contributors to mortality prediction in critically ill patients [47].

The SHAP dependence plots provided additional insight into the nonlinear associations captured by the model. For SBP, very low values were associated with a marked increase in predicted mortality risk, whereas the contribution decreased as SBP approached an intermediate range and showed greater variability at higher values. This pattern may reflect the adverse effects of hypotension and impaired tissue perfusion, while the relationship at higher SBP levels may also be influenced by vasopressor use, preexisting hypertension, or other indicators of illness severity. LDH showed a rapid increase in its contribution to predicted risk at lower-to-moderate elevations, followed by a relative plateau at higher levels. However, the SHAP plot did not identify a single clinically definitive threshold, and LDH should therefore be interpreted as a continuous marker of tissue injury and systemic stress rather than as a binary cutoff [44,45]. Mechanical ventilation showed the greatest contribution to model prediction, probably because it reflects respiratory failure, greater organ dysfunction, and increased treatment intensity. Importantly, mechanical ventilation and several other highly ranked variables may primarily represent the severity of the patient's condition rather than independent prognostic factors. SHAP values explain how individual features contribute to model output but do not establish causal relationships [30,31].

From a clinical implementation perspective, the model is intended for risk assessment once routinely collected clinical data from the first 24 hours after ICU admission are available. Because the predictors are routinely captured in electronic health records, the model could potentially be integrated into electronic health record (EHR) systems to automatically generate individualized estimates of 28-day mortality risk. The predicted probability should be interpreted as a risk stratification tool rather than as a direct treatment recommendation. Patients with higher predicted risk may warrant closer monitoring, earlier clinical reassessment, and greater attention to potential deterioration, whereas lower predicted risk should not be used in isolation to justify de-escalation of care. Therefore, the model is primarily intended to support risk stratification and clinical awareness rather than independently determine treatment decisions or resource allocation. Prospective validation and assessment of calibration and clinical utility are required before implementation in routine practice.

Several methodological limitations should be acknowledged. Some predictors retained in the final model were clinically correlated, such as SBP and mean arterial pressure, PaO_2 and the PaO_2/FiO_2 ratio, and creatinine and blood urea nitrogen (BUN). Although tree-based models are relatively tolerant of correlated predictors in terms of predictive performance, such correlations may distribute feature importance across related variables and affect the stability of SHAP attribution. In addition, the common 16-feature set used for all algorithms was selected using LightGBM and may therefore have favored tree-based models. No formal collinearity-based elimination or model-specific feature-selection procedure was performed. Consequently, SHAP rankings and comparisons between algorithms should be interpreted within the context of the selected feature set.

Another limitation is the potential underestimation of mortality outcomes in the eICU-Sepsis cohort. Because eICU Collaborative Research Database (eICU-CRD) largely relies on outcomes documented during hospitalization and does not provide comprehensive active post-discharge follow-up, deaths occurring after hospital discharge may not have been fully captured. Therefore, the observed 28-day mortality rate in this cohort may be lower than the true mortality rate, which should be considered when interpreting the external validation results.

5. Conclusion

In conclusion, this study developed and externally validated an interpretable XGBoost model for predicting 28-day mortality in ICU patients with sepsis. The model showed comparable discrimination across an internal and two external validation cohorts. SHAP analysis identified several routinely available clinical features that contributed substantially to model prediction. The model may support future risk stratification applications, although prospective validation and clinical impact assessment are required before routine implementation.

Key Points

• This retrospective multicenter cohort study developed and validated an interpretable machine learning model to predict 28-day mortality in adult ICU patients with sepsis.

• The final model used 16 routinely collected variables available by the 24-hour prediction landmark, enabling risk assessment after the first 24 hours of ICU admission.

• Among nine machine learning algorithms, XGBoost showed the best overall predictive performance, with an AUROC of 0.884 (0.871–0.897) in the internal validation cohort.

• External validation of the model in the eICU-Sepsis and DRECC-Sepsis cohorts showed comparable discrimination, with AUROCs of 0.877 (0.872–0.882) and 0.875 (0.856–0.894), respectively.

• SHAP analysis identified mechanical ventilation, Charlson Comorbidity Index, lactate dehydrogenase, age, respiratory rate, and systolic blood pressure as important contributors to predicted 28-day mortality risk.

Availability of Data and Materials

The MIMIC-IV and eICU-CRD databases are publicly available through PhysioNet at https://physionet.org/content/mimiciv/ and https://physionet.org/content/eicu-crd/, respectively, subject to the applicable credentialing, training, and data-use requirements. The DRECC-Sepsis database is not publicly accessible because of patient privacy and institutional data governance restrictions, and no public-access website is available for this dataset; however, data may be made available from the corresponding authors upon reasonable request and subject to appropriate ethical approval.

Author Contributions

ZCL, LHH, PFS, NL and SYZ conceived and designed the study. ZCL, LHH, PFS, CH, JJH, YH and DPL participated in data extraction, data curation, model development, statistical analysis, and manuscript drafting. BYZ and YCH contributed to the conception and design of the study, interpretation of the results, clinical and methodological supervision, and critical revision of the manuscript. NL supervised the project and secured resources. All authors have been involved in revising it critically for important intellectual content. All authors read and approved the final manuscript and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

The study was conducted in accordance with the Declaration of Helsinki. Ethical approval for the DRECC-Sepsis cohort was obtained from the Ethics Committee of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine (Approval No. 2023-964-01). The requirement for written informed consent was waived by the Ethics Committee because this retrospective study used de-identified clinical data. The MIMIC and eICU databases contain de-identified data and were used in accordance with their data-use requirements.

Acknowledgment

Not applicable.

Funding

This work was supported by the Fund of the Zhejiang Key Laboratory of Intelligent Medical Decision Support (IMDS2026Y004).

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

ChatGPT (OpenAI) was used to improve the clarity, grammar, and readability of the manuscript. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the submitted manuscript. After its use, the authors thoroughly reviewed, verified, and revised all content to ensure accuracy and originality.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/BJHM55546.

References

[1] Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet (London, England). 2020; 395: 200–211. https://doi.org/10.1016/S0140-6736(19)32989-7

[2] Mellhammar L, Wollter E, Dahlberg J, Donovan B, Olséen CJ, Wiking PO, et al. Estimating Sepsis Incidence Using Administrative Data and Clinical Medical Record Review. JAMA Network Open. 2023; 6: e2331168. https://doi.org/10.1001/jamanetworkopen.2023.31168

[3] Al Omar S, Alshraideh JA, Oweidat I, Al Qadire M, Khalaf A, Abu Sumaqa Y, et al. Mortality of patients with sepsis in intensive care units at tertiary hospitals in Jordan: Prospective cohort study. Medicine. 2024; 103: e40169. https://doi.org/10.1097/MD.0000000000040169

[4] Ward MA, Kuttab HI, Badgett RG. The Effect of Early Fluid Resuscitation on Mortality in Sepsis: A Systematic Review and Meta-Analysis. Critical Care Medicine. 2025; 53: e1790–e1802. https://doi.org/10.1097/CCM.0000000000006769

[5] Mira JC, Gentile LF, Mathias BJ, Efron PA, Brakenridge SC, Mohr AM, et al. Sepsis Pathophysiology, Chronic Critical Illness, and Persistent Inflammation-Immunosuppression and Catabolism Syndrome. Critical Care Medicine. 2017; 45: 253–262. https://doi.org/10.1097/CCM.0000000000002074

[6] van der Slikke EC, An AY, Hancock REW, Bouma HR. Exploring the pathophysiology of post-sepsis syndrome to identify therapeutic opportunities. EBioMedicine. 2020; 61: 103044. https://doi.org/10.1016/j.ebiom.2020.103044

[7] Chanderraj R, Bartek B, Stringer KA, Tiba MH, Sjoding MW, He Y, et al. Pathogen characteristics are key determinants of distinct host response phenotypes of sepsis. The Journal of Clinical Investigation. 2026; 136: e197346. https://doi.org/10.1172/JCI197346

[8] Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA. 2019; 321: 2003–2017. https://doi.org/10.1001/jama.2019.5791

[9] Singer M, Angus DC, Annane D, Bauer M, Kalil AC, Klompas M, et al. Sepsis. Lancet (London, England). 2026; 407: 1276–1288. https://doi.org/10.1016/S0140-6736(25)02422-5

[10] Qiu X, Lei YP, Zhou RX. SIRS, SOFA, qSOFA, and NEWS in the diagnosis of sepsis and prediction of adverse outcomes: a systematic review and meta-analysis. Expert Review of Anti-

infective Therapy. 2023; 21: 891–900. https://doi.org/10.1080/14787210.2023.2237192
[11] Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, et al. Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit. JAMA. 2017; 317: 290–300. https://doi.org/10.1001/jama.2016.20328
[12] Jiang J, Huang H, Li H, Shen J. Sepsis-related coagulation-inflammation score: Development and validation for predicting 28-day mortality in patients with sepsis. Thrombosis Research. 2025; 255: 109479. https://doi.org/10.1016/j.thromres.2025.109479
[13] Valik JK, Ward L, Tanushi H, Johansson AF, Färnert A, Mogensen ML, et al. Predicting sepsis onset using a machine learned causal probabilistic network algorithm based on electronic health records data. Scientific Reports. 2023; 13: 11760. https://doi.org/10.1038/s41598-023-38858-4
[14] Islam KR, Prithula J, Kumar J, Tan TL, Reaz MBI, Sumon MSI, et al. Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review. Journal of Clinical Medicine. 2023; 12: 5658. https://doi.org/10.3390/jcm12175658
[15] Park JY, Hsu TC, Hu JR, Chen CY, Hsu WT, Lee M, et al. Predicting Sepsis Mortality in a Population-Based National Database: Machine Learning Approach. Journal of Medical Internet Research. 2022; 24: e29982. https://doi.org/10.2196/29982
[16] Wang Y, Gao Z, Zhang Y, Lu Z, Sun F. Early sepsis mortality prediction model based on interpretable machine learning approach: development and validation study. Internal and Emergency Medicine. 2025; 20: 909–918. https://doi.org/10.1007/s11739-024-03732-2
[17] Nikravangolsefid N, Reddy S, Truong HH, Charkviani M, Ninan J, Prokop LJ, et al. Machine learning for predicting mortality in adult critically ill patients with Sepsis: A systematic review. Journal of Critical Care. 2024; 84: 154889. https://doi.org/10.1016/j.jcrc.2024.154889
[18] Mou C, Yang J, Wu Q, Qin L, Lu J. Progress in sepsis prediction models: from traditional scoring systems to multimodal intelligence and clinical translation. Frontiers in Medicine. 2026; 13: 1732164. https://doi.org/10.3389/fmed.2026.1732164
[19] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. [published correction appears in Scientific Data. 2023; 10: 31. https://doi.org/10.1038/s41597-023-01945-2; published correction appears in Scientific Data. 2023; 10: 219. https://doi.org/10.1038/s41597-023-02136-9]. Scientific Data. 2023; 10: 1. https://doi.org/10.1038/s41597-022-01899-x
[20] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data. 2018; 5: 180178. https://doi.org/10.1038/sdata.2018.178
[21] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016; 315: 762–774. https://doi.org/10.1001/jama.2016.0288
[22] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016; 315: 801–810. https://doi.org/10.1001/jama.2016.0287
[23] Gao J, Lu Y, Ashrafi N, Domingo I, Alaei K, Pishgar M. Prediction of sepsis mortality in ICU patients using machine learning methods. BMC Medical Informatics and Decision Making. 2024; 24: 228. https://doi.org/10.1186/s12911-024-02630-z
[24] Zhang SZ, Ding HY, Shen YM, Shao B, Gu YY, Chen QH, et al. Harness machine learning for multiple prognoses prediction in sepsis patients: evidence from the MIMIC-IV database. BMC Medical Informatics and Decision Making. 2025; 25: 152. https://doi.org/10.1186/s12911-025-02976-y
[25] Zhuang J, Huang H, Jiang S, Liang J, Liu Y, Yu X. A generalizable and interpretable model for mortality risk stratification of sepsis patients in intensive care unit. BMC Medical Informatics and Decision Making. 2023; 23: 185. https://doi.org/10.1186/s12911-023-02279-0
[26] Shen L, Wu J, Lan J, Chen C, Wang Y, Li Z. Interpretable machine learning-based prediction of 28-day mortality in ICU patients with sepsis: a multicenter retrospective study. Frontiers in Cellular and Infection Microbiology. 2025; 14: 1500326. https://doi.org/10.3389/fcimb.2024.1500326
[27] Yang M, Chen H, Hu W, Mischi M, Shan C, Li J, et al. Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study. Journal of Medical Internet Research. 2024; 26: e50369. https://doi.org/10.2196/50369
[28] Zhang G, Shao F, Yuan W, Wu J, Qi X, Gao J, et al. Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers. European Journal of Medical Research. 2024; 29: 156. https://doi.org/10.1186/s40001-024-01756-0
[29] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical Care. 2005; 43: 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83
[30] Rodríguez-Pérez R, Bajorath J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. Journal of Medicinal Chemistry. 2020; 63: 8761–8777. https://doi.org/10.1021/acs.jmedchem.9b01101
[31] Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine. 2022; 214: 106584. https://doi.org/10.1016/j.cmpb.2021.106584
[32] Hou N, Li M, He L, Xie B, Wang L, Zhang R, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. Journal of Translational Medicine. 2020; 18: 462. https://doi.org/10.1186/s12967-020-02620-5
[33] Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ (Clinical Research Ed.). 2024; 384: e074819. https://doi.org/10.1136/bmj-2023-074819
[34] Park SW, Yeo NY, Kang S, Ha T, Kim TH, Lee D, et al. Early Prediction of Mortality for Septic Patients Visiting Emergency Room Based on Explainable Machine Learning: A Real-World Multicenter Study. Journal of Korean Medical Science. 2024; 39: e53. https://doi.org/10.3346/jkms.2024.39.e53
[35] Cox EGM, Wiersema R, Eck RJ, Kaufmann T, Granholm A, Vaara ST, et al. External Validation of Mortality Prediction Models for Critical Illness Reveals Preserved Discrimination but Poor Calibration. Critical Care Medicine. 2023; 51: 80–90. https://doi.org/10.1097/CCM.0000000000005712
[36] Campagner A, Agnello L, Carobene A, Padoan A, Del Ben F, Locatelli M, et al. Complete Blood Count and Monocyte Distribution Width-Based Machine Learning Algorithms for Sepsis Detection: Multicentric Development and External Validation Study. Journal of Medical Internet Research. 2025; 27: e55492. https://doi.org/10.2196/55492

[37] Zampieri FG, Mazza B. Mechanical Ventilation in Sepsis: A Reappraisal. Shock (Augusta, Ga.). 2017; 47: 41–46. https://doi.org/10.1097/SHK.0000000000000702
[38] Kim G, Oh DK, Lee SY, Park MH, Lim CM, Korean Sepsis Alliance (KSA) investigators. Impact of the timing of invasive mechanical ventilation in patients with sepsis: a multicenter cohort study. Critical Care (London, England). 2024; 28: 297. https://doi.org/10.1186/s13054-024-05064-1
[39] Yang Y, Yang KS, Hsann YM, Lim V, Ong BC. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. Journal of Critical Care. 2010; 25: 398–405. https://doi.org/10.1016/j.jcrc.2009.09.001
[40] Prescott HC, Harrison DA, Rowan KM, Shankar-Hari M, Wunsch H. Temporal Trends in Mortality of Critically Ill Patients with Sepsis in the United Kingdom, 1988-2019. American Journal of Respiratory and Critical Care Medicine. 2024; 209: 507–516. https://doi.org/10.1164/rccm.202309-1636OC
[41] Fay K, Sapiano MRP, Gokhale R, Dantes R, Thompson N, Katz DE, et al. Assessment of Health Care Exposures and Outcomes in Adult Patients With Sepsis and Septic Shock. JAMA Network Open. 2020; 3: e206004. https://doi.org/10.1001/jamanetworkopen.2020.6004
[42] Martin-Loeches I, Guia MC, Vallecoccia MS, Suarez D, Ibarz M, Irazabal M, et al. Risk factors for mortality in elderly and very elderly critically ill patients with sepsis: a prospective, observational, multicenter cohort study. Annals of Intensive Care. 2019; 9: 26. https://doi.org/10.1186/s13613-019-0495-x
[43] Liu G, Shang Z, Ning N, Li J, Sun W, Fan Y, et al. EVALUATION OF PROGNOSTIC RISK MODELS BASED ON AGE AND COMORBIDITY IN SEPTIC PATIENTS: INSIGHTS FROM MACHINE LEARNING AND TRADITIONAL METHODS IN A LARGE-SCALE, MULTICENTER, RETROSPECTIVE STUDY. Shock (Augusta, Ga.). 2025; 64: 56–64. https://doi.org/10.1097/SHK.0000000000002562
[44] Drent M, Cobben NA, Henderson RF, Wouters EF, van Dieijen-Visser M. Usefulness of lactate dehydrogenase and its isoenzymes as indicators of lung damage or inflammation. The European Respiratory Journal. 1996; 9: 1736–1742. https://doi.org/10.1183/09031936.96.09081736
[45] Qin Y, Wu A, Wang Y, Qin X, Zhang J, Guo X. Serum lactate dehydrogenase level as a predictor of 28-day mortality in critically ill patients with infective endocarditis: a retrospective cohort study from MIMIC IV database. Heart & Lung: the Journal of Critical Care. 2025; 72: 74–82. https://doi.org/10.1016/j.hrtlng.2025.04.002
[46] Guo Q, Li HY, Song WD, Li M, Chen XK, Liu H, et al. Contributions of individual qSOFA elements to assessment of severity and for prediction of mortality. Annals of Medicine. 2024; 56: 2397090. https://doi.org/10.1080/07853890.2024.2397090
[47] Zhou S, Lu Z, Liu Y, Wang M, Zhou W, Cui X, et al. Interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: development and validation. European Journal of Medical Research. 2024; 29: 14. https://doi.org/10.1186/s40001-023-01593-7