

Interpretable Predictive Modeling for Early Sepsis in the Emergency Department: Development and Temporal External Validation

Xuefei Lu

Xinjiang Medical University

Xiaoyu Zhang

Xinjiang Medical University

zhijie Cao

Xinjiang Medical University

Lina Ma

Xinjiang Medical University

Jing Ma

Xinjiang Medical University

Siyi Wang

Xinjiang Medical University

Li Zhang

18099143356@163.com

Xinjiang Medical University

Research Article

Keywords: Sepsis, Prediction model, Emergency department, Machine learning, Interpretability, Early identification, Nursing assessment

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10629603/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose: Existing early warning scores often suffer from low specificity and high false-positive rates in the emergency department (ED). This study aimed to develop an interpretable machine learning model for early sepsis prediction using initial emergency department data.

Methods: In this single-center retrospective study, 1,459 patients with infection or suspected infection were enrolled between January 1 and December 31, 2025. Patients were randomly split into a training set and an internal validation set (7:3 ratio). Eleven predictive variables were selected via univariate analysis, LASSO regression, and the Boruta algorithm. Seven ML algorithms were trained, with hyperparameters optimized using grid search and 10-fold cross-validation. The XGBoost model with optimal performance was selected. To assess real-world generalizability, the final model parameters were frozen and tested on a temporally independent validation cohort from January 1 to March 31, 2026, without any recalibration. Model interpretability was analyzed using SHAP values.

Results: The XGBoost model demonstrated good predictive performance on the internal validation set, with an area under the receiver operating characteristic curve (AUC) of 0.823 (95% confidence interval: 0.783–0.862). In temporal external validation, the model yielded an AUC of 0.972 (95% CI: 0.957–0.985) and a Brier score of 0.073, indicating well-calibrated predictions. Notably, it achieved a specificity of 98.26%, enabling accurate exclusion of low-risk patients and reducing unnecessary clinical alerts. Of the 1459 patients, 347 (23.8%) were diagnosed with sepsis. SHAP analysis identified C-reactive protein (CRP), procalcitonin (PCT), and platelet count (PLT) as the most influential predictors. The model captured nonlinear relationships between key features and sepsis risk. SHAP waterfall plots quantified individual prediction risk by illustrating the contribution of specific features, offering actionable insights for clinical assessment.

Conclusion: We developed an interpretable ML model for the early prediction of sepsis in the ED. The model provides not only accurate risk stratification but also transparent, feature-based explanations for its predictions, which can guide nursing assessment and decision-making. It represents a promising and transparent tool with potential for clinical translation.

Introduction

Sepsis is a life-threatening syndrome characterized by organ dysfunction caused by a dysregulated host response to infection^[1]. It represents a major global public health challenge. According to the World Health Organization (WHO) Global Burden of Disease report, there are approximately 48.9 million new cases and 11 million deaths attributed to sepsis annually worldwide, accounting for 19.7% of all global deaths^[2]. The emergency department (ED) serves as the critical frontline for sepsis care, as statistics indicate that approximately 68.7% of sepsis patients initially present there, making it the pivotal site for early identification and intervention^[2, 3]. Both national and international guidelines emphasize that timely management during this critical “golden window” in the ED can significantly reduce sepsis mortality^[4, 5]. However, early diagnosis remains difficult; studies report that for 30–70% of sepsis patients, the causative pathogen and its source cannot be determined early in the clinical course, complicating timely recognition^[6]. Delays in identification and treatment can precipitate rapid progression to septic shock and multiple organ dysfunction. Evidence suggests that each hour of delay in administering appropriate antibiotics is associated with a 7.6% increase in mortality^[7]; delays in treatment exceeding 3 hours elevate the risk of complications such as acute kidney injury and acute respiratory distress syndrome^[8], potentially leading to irreversible tissue damage^[9]. Therefore, implementing efficient screening tools to facilitate rapid intervention in the ED is crucial.

Currently, clinical practice primarily relies on the Quick Sequential Organ Failure Assessment (qSOFA) and Sequential Organ Failure Assessment (SOFA) scoring systems for early risk stratification. However, their utility in the ED is constrained by well-documented limitations. While qSOFA is simple and easy to use, its low sensitivity in the ED population with infections leads to a high rate of missed diagnoses in early-stage sepsis. Conversely, the SOFA score, though more comprehensive, is predominantly validated and utilized in the intensive care unit (ICU). Its application is less practical for initial screening in the ED due to the complexity of the score and the dynamic, time-pressured environment where comprehensive data may be initially limited^[10–12]. Furthermore, these scores are inherently limited by their basis in linear combinations of a few variables, which are inadequate for modeling the complex, nonlinear pathophysiological processes that occur in early sepsis. Thus, within the information-constrained and high-acuity context of the ED, the early warning performance of these conventional tools remains suboptimal and necessitates improvement.

In recent years, machine learning (ML) has demonstrated significant potential for analyzing healthcare big data. Compared to traditional scoring systems, ML algorithms offer a superior capacity to integrate and analyze multidimensional, non-linear clinical data—such as vital signs, laboratory parameters, and demographic characteristics—generated in the early hours after a patient’s presentation. They can automatically uncover complex interaction patterns and underlying trends^[13], thereby enabling the development of more sensitive and specific early prediction models. Consistent with this, studies have shown that ML-based predictive models may outperform conventional methods in identifying high-risk patients^[14, 15]. However, a critical gap exists in translating this potential to emergency department (ED) sepsis

care. Most existing ML-based sepsis prediction models are developed using comprehensive data from the intensive care unit (ICU) or from a relatively late post-admission period^[16-19]. Furthermore, many high-performance models function as "black boxes," lacking clinical interpretability due to their structural complexity^[20,21], which severely limits their utility in the rapid, high-stakes decision-making environment of the ED.

Therefore, for the specific population of patients with suspected infection in the ED, particularly in regions like Northwest China, there is a crucial need to construct an early warning model that combines high predictive performance with clinical interpretability, using data collected within the first 6 hours after admission. Such a tool is essential for achieving the goal of "early detection and early intervention." Currently, studies fulfilling these criteria remain scarce.

To address these shortcomings, this study aims to construct and validate an early-risk prediction model for sepsis tailored to the ED setting. We conducted a retrospective analysis of clinical data from 1,459 patients with infection or suspected infection admitted to the ED of a tertiary teaching hospital in China between January 1 and December 31, 2025. By applying and comparing multiple machine learning algorithms, our goal is to provide clinicians with a more accurate, earlier, and interpretable decision-support tool to facilitate the early warning and intervention of sepsis.

1 Clinical Data and Methods

1.1 Study Design and Population

This retrospective cohort study consecutively enrolled 1,459 patients with confirmed or suspected infections who presented to the Emergency Department of XX between January 1, 2025, and December 31, 2025.

Inclusion criteria were: 1) Patients meeting the diagnostic criteria for infection or suspected infection; 2) Age ≥ 18 years.

Exclusion criteria were: 1) Age < 18 years; 2) Severe missing clinical data; 3) Patients with an emergency observation time < 24 hours who were not diagnosed with sepsis; 4) Pregnant or lactating women; 5) Patients with advanced malignant tumors, severe immunodeficiency diseases, or in a terminal condition.

Based on the definitions and diagnostic criteria outlined in the "Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock 2026"^[22], patients were classified into a sepsis group ($n = 347$) and a non-sepsis group ($n = 1112$). The sample size was determined by data availability from the electronic medical records during the study period. To ensure model reliability, we adhered to the Events Per Variable (EPV) principle. The EPV ratio was approximately 31:1 (347 sepsis events / 11 predictors), substantially exceeding the recommended minimum threshold of 10:1 for developing robust prediction models.

Ethical approval was obtained from the Institutional Review Board of XX (Approval No. XX). Given the retrospective design utilizing anonymized data without impacting clinical management, the requirement for informed consent was waived.

1.2 Assessment and Diagnostic Criteria

Diagnostic criteria for infection or suspected infection^[23-27]: ☐ Abnormal body temperature ($>38^{\circ}\text{C}$ or $<36^{\circ}\text{C}$) within 72 hours of admission or presentation; ☐ Abnormal white blood cell count ($>12 \times 10^9/\text{L}$ or $<4 \times 10^9/\text{L}$); ☐ C-reactive protein (CRP) levels exceeding 2 standard deviations above the normal range; ☐ Procalcitonin exceeding the normal range by 2 standard deviations; ☐ Clinical manifestations or imaging evidence of a definite or suspected site of infection.

Definition: Patients meeting any two of items ☐, ☐, and ☐, plus any one of criteria ☐ and ☐, were defined as having a "definite infection." Patients meeting only any one of criteria ☐, ☐, and ☐, along with criteria ☐, were defined as having a "suspected infection."

The diagnosis of sepsis was based on the "Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock 2026"^[22], which adheres to the Sepsis-3 definition. Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection^[1]. Operationally, sepsis is diagnosed in the presence of a confirmed or suspected infection, accompanied by an acute increase in the Sequential Organ Failure Assessment (SOFA) score of ≥ 2 points from baseline. Participants meeting these criteria were included in the sepsis group; those not meeting them were included in the non-sepsis group.

1.3 Data Collection

Two trained investigators independently extracted data from the hospital's electronic medical record system. Collected variables included demographic characteristics, first vital signs recorded upon ED arrival, initial laboratory test results, and comorbidities.

To ensure data quality, variables with missing values exceeding 30% were excluded from the analysis. For variables with $\leq 30\%$ missing data, values were imputed using multiple imputation methods. Subsequently, the derivation cohort was randomly divided into a training set and an internal validation set in a 7:3 ratio. All continuous variables (e.g., age, vital signs, and laboratory indices) were entered as raw, ungrouped values without arbitrary cutoffs, thereby preserving statistical power and avoiding subjective information loss.

Within the training set, feature selection was performed using a three-step approach: univariate analysis, least absolute shrinkage and selection operator (LASSO) regression, and the Boruta algorithm. The final core feature set for model construction was defined as the intersection of variables identified by all three methods. The study data were collected from January to December 2025, with a median observation window of 72 hours (IQR: 48–96 hours) following ED arrival. A temporally independent external validation cohort (January 1 – March 31, 2026) was constructed using identical eligibility criteria and variable definitions. This cohort was entirely withheld from model building and feature selection.

As a single-center retrospective study, inherent selection bias and information bias are unavoidable. To mitigate these biases, we strictly adhered to predefined inclusion and exclusion criteria during data extraction. All clinical and laboratory assessments followed standardized hospital protocols.

1.4 Construction of Machine Learning Models

Using the selected feature set, prediction models were developed with seven machine learning algorithms: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), and Artificial Neural Network (ANN). For each algorithm, hyperparameters were optimized via grid search combined with 10-fold cross-validation to identify the configuration yielding the best predictive performance.

1.5 Model Evaluation and Selection of the Optimal Model

All seven models were evaluated and compared on the internal validation set. Discriminative ability was primarily assessed by the area under the receiver operating characteristic curve (AUC), supplemented by metrics including accuracy, sensitivity, specificity, precision, and F1-score. Calibration, reflecting the agreement between predicted probabilities and observed outcomes, was evaluated using calibration curves and the Brier score. The clinical utility of the models across different probability thresholds was analyzed with decision curve analysis (DCA).

The optimal model for subsequent interpretation was selected based on a comprehensive assessment of its discriminative power, calibration, and clinical net benefit as measured by the above metrics. Once the final model was selected, all parameters were frozen and directly applied to the temporally independent external validation cohort without any recalibration, ensuring an unbiased assessment of model generalizability.

1.6 Model Interpretability Analysis

To render the optimal model clinically interpretable, we applied the SHapley Additive exPlanations (SHAP) framework. Using the training set data, SHAP values were calculated to quantify the contribution of each feature to individual predictions. Interpretation was conducted at both the global and local levels through multiple visualizations. Globally, SHAP summary plots (beeswarm plots) and feature importance plots were generated to illustrate the overall ranking and directional effect (positive or negative) of all features. To elucidate nonlinear associations, SHAP dependence plots were used for key features. Locally, for individual patient predictions, SHAP waterfall plots and force plots were employed to visually decompose and attribute the prediction to the contribution of each specific feature value. The same SHAP interpretability framework was maintained for the external cohort, enabling individualized attribution analysis of representative high- and low-risk cases to verify the stability of the model's decision logic beyond the development data.

1.7 Statistical Analysis

All analyses were performed using R software. Univariate analysis and statistical comparisons utilized R (version 4.3.0), while machine learning modeling employed R (version 4.5.1) and Python (version 3.10.4). The normality of continuous variables was assessed using the Shapiro-Wilk test. Normally distributed variables are presented as the mean $\pm$ standard deviation and were compared using the independent samples t-test. Non-normally distributed variables are presented as the median (interquartile range) and were compared using the Mann-Whitney U test. Categorical variables are expressed as frequency (percentage) and were compared using the chi-square test or Fisher's exact test when the expected cell count was less than 5. A two-sided p-value < 0.05 was considered statistically significant for all tests.

2 Results

2.1 Univariate Analysis of Baseline Characteristics and Predictive Factors

A total of 1,459 patients with confirmed or suspected infection were included in the final cohort, comprising 347 patients (23.8%) in the sepsis group and 1,112 patients (76.2%) in the non-sepsis group, consistent with the predefined inclusion and exclusion criteria. The flowchart of participant selection is shown in Figure 1.

Patients were then randomly allocated to a training set (n = 1,022) or a validation set (n = 437) in a 7:3 ratio. The training set contained 243 sepsis patients (23.8%) and 779 non-sepsis group (76.2%). The baseline characteristics of patients in the training set are detailed in Table 1.

Table 1. Baseline Characteristics of the Training Cohort (n=1022)

Variables	Total (n = 1022)	0 (n = 779)	1 (n = 243)	Statistic	P
Male, n(%)				$\chi^2=1.112$	0.292
0	354 (34.638)	263 (33.761)	91 (37.449)		
1	668 (65.362)	516 (66.239)	152 (62.551)		
Hypertension, n(%)				$\chi^2=1.745$	0.187
0	580 (56.751)	451 (57.895)	129 (53.086)		
1	442 (43.249)	328 (42.105)	114 (46.914)		
Diabetes Mellitus, n(%)				$\chi^2=10.187$	0.001
0	761 (74.462)	599 (76.893)	162 (66.667)		
1	261 (25.538)	180 (23.107)	81 (33.333)		
Coronary Heart Disease, n(%)				$\chi^2=0.321$	0.571
0	659 (64.481)	506 (64.955)	153 (62.963)		
1	363 (35.519)	273 (35.045)	90 (37.037)		
Cerebral Infarction, n(%)				$\chi^2=1.067$	0.302
0	752 (73.581)	567 (72.786)	185 (76.132)		
1	270 (26.419)	212 (27.214)	58 (23.868)		
Initial lactate remeasurement, n (%)				$\chi^2=2.920$	0.571
0	605 (59.198)	465 (59.692)	140 (57.613)		
1	365 (35.714)	271 (34.788)	94 (38.683)		
2	25 (2.446)	22 (2.824)	3 (1.235)		
3	19 (1.859)	15 (1.926)	4 (1.646)		
4	8 (0.783)	6 (0.770)	2 (0.823)		
Vasopressors and inotropic agents delay time, n(%)				$\chi^2=18.215$	0.001
0	353 (34.540)	293 (37.612)	60 (24.691)		
1	448 (43.836)	326 (41.849)	122 (50.206)		
2	153 (14.971)	113 (14.506)	40 (16.461)		
3	48 (4.697)	30 (3.851)	18 (7.407)		
4	20 (1.957)	17 (2.182)	3 (1.235)		
Culture isolation delay time, n(%)				$\chi^2=12.897$	0.012
0	175 (17.123)	144 (18.485)	31 (12.757)		
1	606 (59.295)	462 (59.307)	144 (59.259)		
2	182 (17.808)	137 (17.587)	45 (18.519)		
3	33 (3.229)	18 (2.311)	15 (6.173)		
4	26 (2.544)	18 (2.311)	8 (3.292)		

Urinary catheterization, n(%)				$\chi^2=0.215$	0.643
0	53 (5.186)	39 (5.006)	14 (5.761)		
1	969 (94.814)	740 (94.994)	229 (94.239)		
Mechanical ventilation, n(%)				$\chi^2=9.361$	0.002
0	135 (13.209)	117 (15.019)	18 (7.407)		
1	887 (86.791)	662 (84.981)	225 (92.593)		
Sedative, n(%)				$\chi^2=0.135$	0.713
0	511 (50.000)	387 (49.679)	124 (51.029)		
1	511 (50.000)	392 (50.321)	119 (48.971)		
Analgesic drugs, n(%)				$\chi^2=0.932$	0.334
0	448 (43.836)	348 (44.673)	100 (41.152)		
1	574 (56.164)	431 (55.327)	143 (58.848)		
CRRT implementation, n(%)				$\chi^2=4.136$	0.042
0	922 (90.215)	711 (91.271)	211 (86.831)		
1	100 (9.785)	68 (8.729)	32 (13.169)		
Multidrug-resistant bacterial infection, n (%)				$\chi^2=0.165$	0.684
0	683 (66.830)	518 (66.496)	165 (67.901)		
1	339 (33.170)	261 (33.504)	78 (32.099)		
Antibiotic delay time, n(%)				$\chi^2=22.586$	<.001
0	24 (2.348)	20 (2.567)	4 (1.646)		
1	578 (56.556)	449 (57.638)	129 (53.086)		
2	290 (28.376)	221 (28.370)	69 (28.395)		
3	84 (8.219)	48 (6.162)	36 (14.815)		
4	44 (4.305)	39 (5.006)	5 (2.058)		
5	2 (0.196)	2 (0.257)	0 (0.00)		
Interleukin 6, Mean $\pm$ SD	378.713 $\pm$ 1073.072	372.773 $\pm$ 1057.371	397.525 $\pm$ 1123.556	t = -0.303	0.762
Age, M (Q ₁ , Q ₃)	62.000 (51.000, 75.000)	62.000 (51.000, 75.000)	62.000 (54.000, 74.000)	Z = -0.517	0.605
Weight, M (Q ₁ , Q ₃)	65.000 (60.000, 75.000)	65.000 (60.000, 75.000)	65.000 (59.500, 73.000)	Z = -0.758	0.448
Temperature, M (Q ₁ , Q ₃)	36.900 (36.600, 37.500)	36.900 (36.500, 37.500)	37.000 (36.600, 37.500)	Z = -1.248	0.212
Heart Rate, M (Q ₁ , Q ₃)	88.000 (75.000, 104.000)	86.000 (72.500, 102.000)	92.000 (77.000, 110.000)	Z=-2.767	0.006
Respiratory Rate, M (Q ₁ , Q ₃)	20.000 (19.000, 20.000)	20.000 (18.000, 20.000)	20.000 (19.000, 20.000)	Z = -0.366	0.714
Systolic Blood Pressure, M (Q ₁ ,	124.000 (116.000, 141.000)	125.000 (119.000, 142.000)	120.000 (109.500, 133.000)	Z = -4.098	<.001

Q ₃)					
Diastolic Blood Pressure, M (Q ₁ , Q ₃)	76.000 (66.000, 80.000)	76.000 (66.000, 80.000)	75.000 (66.000, 80.000)	Z=-0.539	0.590
GCS, M (Q ₁ , Q ₃)	15.000 (11.000, 15.000)	15.000 (12.000, 15.000)	15.000 (11.000, 15.000)	Z = -0.804	0.421
APACHE II Score, M (Q ₁ , Q ₃)	16.000 (15.000, 19.000)	16.000 (15.000, 19.000)	16.000 (15.000, 19.000)	Z = -0.225	0.822
VTE, M (Q ₁ , Q ₃)	5.000 (5.000, 6.000)	5.000 (5.000, 6.000)	5.000 (5.000, 6.000)	Z = -0.848	0.396
CRP, M (Q ₁ , Q ₃)	49.900 (17.000, 91.000)	42.000 (14.300, 83.850)	78.900 (42.550, 91.000)	Z = -7.013	<.001
PCT, M (Q ₁ , Q ₃)	0.295 (0.085, 2.320)	0.220 (0.070, 1.190)	1.540 (0.190, 12.330)	Z = -8.652	<.001
Lac, M (Q ₁ , Q ₃)	1.700 (1.300, 2.500)	1.700 (1.300, 2.490)	1.700 (1.200, 2.600)	Z=-0.160	0.873
D dimer1, M (Q ₁ , Q ₃)	856.000 (327.000, 2342.500)	773.000 (261.000, 2343.000)	1183.000 (532.000, 2326.000)	Z = -3.073	0.002
WBC, M (Q ₁ , Q ₃)	9.350 (6.950, 14.272)	9.500 (7.000, 14.045)	8.940 (6.080, 14.960)	Z = -0.213	0.831
PLT, M (Q ₁ , Q ₃)	199.000 (127.000, 257.000)	207.000 (139.000, 261.000)	166.000 (90.500, 238.000)	Z=-5.078	<.001
Cr, M (Q ₁ , Q ₃)	70.970 (50.630, 115.840)	69.690 (48.840, 110.600)	83.290 (54.525, 142.560)	Z=-3.452	<.001
Total Bilirubin, M (Q ₁ , Q ₃)	17.580 (12.620, 26.088)	16.940 (12.410, 25.585)	19.530 (13.990, 27.285)	Z=-2.622	0.009
X6-hour average urine output, M (Q ₁ , Q ₃)	58.000 (53.000, 63.000)	59.000 (55.000, 64.000)	56.000 (50.000, 60.000)	Z = -4.262	<.001

Note: Continuous variables are presented as the median (Q1, Q3) or mean $\pm$ SD. Categorical variables are presented as n (%). t: t-test, Z: Mann-Whitney test, χ^2 : Chi-square test. Explanation of Categorical Variable Coding: Binary variables (e.g., gender, comorbidities) are coded as 0 = No, 1 = Yes; delay variables are coded as 0 = No delay, 1 = Delay $\leq$ 1 hour, 2 = Delay $\leq$ 2 hours, 3 = Delay $\leq$ 3 hours, 4 = Delay $\leq$ 4 hours, 5 = Delay $>$ 4 hours.

Data are presented as median (IQR) or n (%). Missing values were observed for 15 variables, all with rates $<1\%$: PCT (n=7, 0.48%), temperature (n=7, 0.48%), vasopressor-inotrope delay time (n=6, 0.41%), diastolic blood pressure (n=5, 0.34%), sedative (n=4, 0.27%), multidrug-resistant bacterial infection (n=3, 0.21%), male (n=2, 0.14%), APACHE II score (n=2, 0.14%), D-dimer (n=2, 0.14%), CRP (n=2, 0.14%), Cr (n=2, 0.14%), Total bilirubin (n=1, 0.07%), Lac (n=1, 0.07%), WBC (n=1, 0.07%), and PLT (n=1, 0.07%). All other variables had no missing data. Missing values were handled using multiple imputation prior to analysis.

2.2 Selection of Core Predictive Variables

Candidate predictors were screened using three independent methods: univariate analysis, the least absolute shrinkage and selection operator (LASSO) regression, and the Boruta algorithm. The definitive set of core variables for model construction was derived from the intersection of features selected by all three methods. In the training set, univariate analysis identified 15 variables with statistically significant differences ($P < 0.05$) between the sepsis group and the non-sepsis group. LASSO regression, with the optimal regularization parameter ($\lambda_{1se} = 0.021$) determined by 10-fold cross-validation, selected 20 variables with non-zero coefficients (Figure 2A, B). The Boruta algorithm, based on comparative feature importance, identified 28 relevant variables (Figure 2C, D). A Venn diagram illustrates the overlap between the variable sets from each method (15, 20, and 28 variables, respectively) (Figure 2E). Their intersection yielded 11 core variables retained for predictive modeling: history of diabetes, delay in vasopressor administration, delay in culture collection, use of mechanical ventilation, systolic blood pressure, C-reactive protein, procalcitonin, D-dimer, platelet count, total bilirubin, and 6-hour mean urine output.

2.3 Model Construction and Performance Comparison

Using the 11 selected core variables, prediction models were developed with seven ML algorithms: logistic regression (LR), decision trees (DT), random forests (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), support vector machines (SVM), and artificial neural networks (ANN). The performance metrics of all models on both the training and internal validation sets are detailed in Table 2.

All models exhibited strong discriminative ability on the training set, with the RF and LightGBM models achieving the highest AUC values of 0.990. Performance remained robust on the internal validation set without signs of substantial overfitting. Among them, the LightGBM model demonstrated the best performance with an AUC of 0.823 (95% CI: 0.783–0.862). It was closely followed by XGBoost (AUC=0.822, 95% CI: 0.784–0.862) and RF (AUC=0.806, 95% CI: 0.761–0.848). The DT and LR models showed moderate performance, with AUC values of 0.719 (95% CI: 0.666–0.771) and 0.702 (95% CI: 0.647–0.753), respectively. The ANN model demonstrated the lowest discriminative ability in this cohort, with an AUC of 0.661 (95% CI: 0.603–0.718).

Table 2. Performance of Machine Learning Models in Training and Internal Validation

	Model	Logistic	Decision Tree	Random Forest	XGBoost	LightGBM	SVM	ANN
Training set	AUC	0.740	0.920	0.990	0.980	0.990	0.950	0.750
	AUC 95% CI Lower	0.710	0.900	0.990	0.970	0.990	0.930	0.720
	AUC 95% CI Upper	0.780	0.940	0.990	0.990	0.990	0.970	0.780
	Accuracy	0.780	0.850	0.950	0.910	0.950	0.940	0.770
	Precision	0.610	0.700	0.910	0.920	0.900	0.930	0.620
	Sensitivity	0.210	0.680	0.880	0.670	0.890	0.810	0.120
	Specificity	0.960	0.910	0.970	0.980	0.970	0.980	0.980
	F1 Score	0.310	0.690	0.900	0.770	0.890	0.860	0.200
	Kappa	0.220	0.590	0.870	0.710	0.860	0.820	0.130
	Youden's J	0.170	0.590	0.860	0.650	0.860	0.790	0.100
	PPV	0.610	0.700	0.910	0.920	0.900	0.930	0.620
	NPV	0.800	0.900	0.960	0.900	0.970	0.940	0.780
Internal validation set	AUC	0.702	0.719	0.806	0.822	0.823	0.736	0.661
	AUC 95% CI Lower	0.647	0.666	0.761	0.784	0.783	0.679	0.603
	AUC 95% CI Upper	0.753	0.771	0.848	0.862	0.862	0.791	0.718
	Accuracy	0.757	0.721	0.792	0.778	0.778	0.792	0.753
	Precision	0.475	0.417	0.573	0.548	0.535	0.597	0.423
	Sensitivity	0.183	0.433	0.490	0.385	0.510	0.385	0.106
	Specificity	0.937	0.811	0.886	0.901	0.862	0.919	0.955
	F1 Score	0.264	0.425	0.528	0.452	0.522	0.468	0.169
	Kappa	0.152	0.240	0.396	0.318	0.378	0.346	0.082
	Youden's J	0.120	0.244	0.376	0.286	0.371	0.304	0.061
	PPV	0.475	0.417	0.573	0.548	0.535	0.597	0.423
	NPV	0.786	0.821	0.848	0.824	0.849	0.827	0.774

Note: Performance metrics of seven machine learning models (LR, DT, RF, XGBoost, LightGBM, SVM, ANN) evaluated on the training and internal validation cohorts. AUC = Area Under the Curve; CI = Confidence Interval; PPV = Positive Predictive Value; NPV = Negative Predictive Value.

2.4 Validation of the Machine Learning–Based Sepsis Prediction Model

2.4.1 Internal validation

The seven developed prediction models were comprehensively evaluated and compared on the internal validation set across three key dimensions: discrimination, calibration, and clinical utility (Figure 3).

Regarding discrimination, the ROC curves (Figure 3A) confirmed that all models possessed discriminative ability. Among them, the LightGBM model achieved the highest AUC value of 0.823 (95% CI: 0.783–0.862); closely followed by the XGBoost model with an AUC of 0.822 (95% CI:

0.784–0.862). The performance of these two models was comparable and superior to that of the remaining models. In terms of calibration, the calibration curve (Figure 3B) revealed that the predicted probabilities of the XGBoost model aligned most closely with the observed outcomes, supported by the lowest Brier score of 0.142 (95% CI: 0.123–0.163), indicating the best calibration. Finally, decision curve analysis (DCA) (Figure 3C) demonstrated that the XGBoost model provided the highest clinical net benefit across a wide range of probability thresholds.

In summary, while the LightGBM model had a marginally higher AUC, the XGBoost model demonstrated superior and more balanced performance in calibration and clinical utility. Therefore, considering the paramount importance of predictive accuracy and reliability for clinical application, the XGBoost model was selected as the optimal predictive model for subsequent interpretability analysis. The model was then frozen and advanced to temporal external validation and SHAP-based interpretation, without further tuning.

2.4.2 Temporal external validation

To evaluate the generalizability and robustness of the XGBoost model on completely independent data, we adopted a temporal external validation strategy. After completion of model development, all parameters were frozen and no further retraining or tuning was performed. The locked model was directly applied to an independent patient cohort from the same institution spanning January 1 to March 31, 2026, simulating a forward-in-time deployment scenario.

As shown in Figure 4, the model retained exceptional performance in the temporal validation set. The receiver operating characteristic curve (Figure 4A) demonstrated strong discrimination, with an area under the curve (AUC) of 0.972 (95% CI: 0.957–0.985). Calibration analysis (Figure 4B) revealed close agreement between predicted probabilities and observed sepsis incidence, supported by a low Brier score of 0.073, indicating well-calibrated risk estimates. Decision curve analysis (Figure 4C) further confirmed that the XGBoost model yielded the highest net clinical benefit across a broad threshold probability range of approximately 10% to 80%, outperforming both the “treat-all” and “treat-none” strategies. The confusion matrix (Figure 4D) showed that the model achieved a specificity of 98.26% (corresponding to a very low false-positive rate), accurately excluding low-risk individuals. Combined with the DCA results, this suggests the model can effectively support risk stratification while minimizing unnecessary clinical alerts.

In summary, the XGBoost model maintained excellent performance when transferred to a temporally independent cohort, confirming its robust generalizability and practical translational value for early sepsis screening in the emergency department.

2.5 Model Interpretability Analysis

To elucidate the decision logic of the optimal (XGBoost) model and enhance its clinical trustworthiness, we applied the SHapley Additive exPlanations (SHAP) framework for a comprehensive interpretability analysis, quantifying each feature's contribution to individual predictions at both global and local levels.

The results are presented in Figure 5. The SHAP feature importance bar chart (Figure 5A) provides a global ranking, identifying C-reactive protein (CRP), procalcitonin (PCT), and platelet count (PLT) as the three most influential predictors. The SHAP summary plot (Figure 5B) further illustrates the directional relationship between feature values and the model's output: elevated CRP levels (red dots) are associated with increased predicted risk (positive SHAP values), whereas a higher 6-hour mean urine output (blue dots on the left) correlates with decreased risk (negative SHAP values).

To further elucidate the model's decision-making logic for specific high-risk individuals, we generated a SHAP waterfall plot for the corresponding high-risk sample. As shown in Figure 5C, when the model rendered a high-risk prediction, the extremely elevated C-reactive protein (CRP) level (+0.07) served as the dominant driver, supplemented by increased procalcitonin (PCT, +0.04) and platelets (PLT) within the normal range (+0.03), which collectively propelled the predicted value upward. Meanwhile, the elevated D-dimer level (-0.01, blue dot) exerted a minor negative regulatory effect on the prediction, attempting to attenuate the risk probability but failing to offset the strong positive contributions of the core inflammatory indicators. This interpretability analysis demonstrates that the decision-making logic of the XGBoost model aligns closely with the pathophysiological mechanisms underlying severe infection and organ dysfunction secondary to inflammation.

In contrast, to validate the model's discriminative logic in low-risk populations, Figure 4D illustrates the decision-making process for a representative low-risk sample. The base value ($E[f(X)]$) for this sample was 0.234, whereas the model's final predicted risk was markedly lower at $f(x)=0.04$. As indicated by the prominent blue bars, critically elevated levels of CRP (-0.07), PCT (-0.04), and PLT (-0.03)—key inflammatory and hematological parameters—exerted substantial negative regulatory effects on the prediction. Furthermore, physiological metrics such as total bilirubin and 6-hour average urine output also contributed negatively (each approximately -0.01). This collective downward modulation successfully counteracted the baseline risk, suppressing the final probability to a minimal level. These

findings indicate that even in complex cases characterized by moderate baseline risk yet pronounced inflammatory marker elevations, the model can sensitively capture nuanced feature interactions to render clinically

congruent low-risk predictions, underscoring its robust and comprehensive discriminative capacity.

To elucidate the nonlinear associations between key predictors and the model outputs, SHAP dependence plots were generated for all 11 variables (Figure 6). As illustrated, the relationships between key inflammatory markers (CRP, PCT) and the predicted outcome (SHAP value) are inherently nonlinear. The SHAP value for PCT increases steeply at lower concentrations (0–10 ng/mL) before plateauing, whereas the contribution of CRP peaks at moderate-to-high levels (approximately 50–80 mg/L) and subsequently declines. Other indicators, such as platelet count (PLT) and urine output, also exhibit distinct, non-monotonic patterns. Collectively, these plots demonstrate the model's capacity to capture the complex, nonlinear interactions among clinical biomarkers.

3 Discussion

This study developed and internally validated an XGBoost model for the early prediction of sepsis, using clinical data obtained within 6 hours of Emergency Department presentation. The model demonstrated strong discriminatory ability, good calibration, and a favorable clinical net benefit in the internal validation cohort. Crucially, by applying SHAP explainability analysis, we elucidated the key drivers of the model's predictions and delineated their complex, non-linear relationships with sepsis risk, thereby enhancing the model's transparency and potential utility as a clinical decision-support tool.

In contrast to most previous studies predominantly using ICU or post-admission data^[28–30], this study advances the prediction window to the critical first hours of emergency care, enabling earlier detection and potential intervention. Although machine learning has been employed for sepsis prediction in prior work, the interpretability of these models has often been neglected^[28, 31–33]. Our model not only achieved robust predictive performance in the internal validation set (AUC: 0.823), but also, through the SHAP framework, provided both global and local explanations for its predictions. This directly addresses the key challenge of transforming complex "black-box" models into transparent, clinically actionable decision-support tools.

The model's predictive strength stems from its capacity to capture complex, non-linear pathophysiological relationships. SHAP dependence plots reveal that key biomarkers, such as PCT and CRP, do not exhibit simple linear correlations with sepsis risk but demonstrate distinct threshold and saturation dynamics. This finding carries significant implications for clinical practice, particularly nursing assessment. It suggests a shift in focus from evaluating whether a single absolute value crosses a static threshold to interpreting the clinical significance of dynamic trends. For instance, a rapid rise in PCT within the normal range may constitute an early warning signal as critical as a definitively elevated level. Furthermore, the model attributes high importance to integrated, multi-system indicators spanning inflammation, coagulation, and organ perfusion. This provides a systematic, data-driven framework for emergency nursing assessment, reinforcing the holistic clinical principle of synthesizing multi-dimensional information for patient evaluation during rapid triage.

The core clinical contribution of this study lies in its integrated "prediction-with-interpretation" output. SHAP waterfall plots deconstruct an individual's predicted risk into the contribution of specific features; for instance, clarifying that a particular high-risk alert is primarily driven by "a sharp rise in CRP" coupled with "a progressive decline in urine output." This interpretability is key to its potential integration into clinical workflows. During nursing shift handovers or patient assessments, these explanations can direct attention to specific, high-stakes clinical dimensions, enabling more targeted monitoring, documentation, and precise communication with physicians. This represents a paradigm shift in decision support—from delivering an opaque "risk score" to providing transparent, "evidence-based assessment cues" that guide clinical reasoning and action.

This study has several limitations that warrant consideration. First, while we incorporated an independent 2026 temporal validation cohort to preliminarily assess out-of-sample stability, this work remains fundamentally a single-center retrospective analysis. Local care pathways and case-mix characteristics may constrain broader generalizability, which should be further confirmed in prospective, multi-center external validation studies. Second, all predictors were derived from the initial 6-hour static window after ED arrival, capturing only a single snapshot rather than the dynamic evolution of physiological and laboratory trajectories over longer time horizons. Third, the model relied exclusively on structured numeric data and did not harness unstructured nursing documentation, such as clinician-recorded observations of lethargy, mottled skin, or peripheral hypoperfusion, which often carry early warning signals. This omission may have limited the model's sensitivity to subtle deterioration and reduced its bedside fidelity. Future research should focus on external validation across diverse settings and explore integrating natural language processing to extract insights from nursing documentation, thereby building a next-generation, nursing-aware sepsis early-warning system.

4 Conclusion

In conclusion, this study developed and internally validated an interpretable machine learning model for the early prediction of sepsis in the emergency department. Beyond its function as a predictive tool, the model—through explainable outputs—translates complex data patterns into clinically actionable insights. This advances sepsis care by providing a transparent, evidence-based framework for risk assessment, supporting a shift toward more proactive and precise clinical management.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Review Committee of the First Affiliated Hospital of Xinjiang Medical University (Approval No. K202511-25). As a retrospective study using anonymized electronic medical record data, the requirement for informed consent was waived by the ethics committee in accordance with national regulations and institutional policies. All procedures complied with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable. This study does not contain any individual person's identifiable images, videos, or private information that requires separate consent for publication.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy regulations imposed by the ethics committee. Anonymized aggregated data and model code are available from the first author, Xuefei Lu, upon reasonable request and with formal permission from the First Affiliated Hospital of Xinjiang Medical University.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Innovation Team Cultivation Project of the First Affiliated Hospital of Xinjiang Medical University (Critical and Emergency Care Innovation Team: Promotion and Application of Key Technologies for Sepsis Early Warning and Treatment, Grant No. cxt202414). The funding body had no role in study design, data collection, analysis, decision to publish, or manuscript preparation.

Authors' contributions

Conceptualization, supervision, project administration: Li Zhang; Data curation, investigation: Xuefei Lu, Xiaoyu Zhang, Zhijie Cao, Lina Ma, Jing Ma, Siyi Wang; Methodology, software, formal analysis: Xuefei Lu, Xiaoyu Zhang; Writing – original draft: Xuefei Lu; Writing – review & editing: All authors. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the staff of the Emergency Department and the Information Center of the First Affiliated Hospital of Xinjiang Medical University for their assistance in data extraction and management. We also acknowledge the nursing team whose bedside documentation forms the basis of clinical context in this study.

Authors' information

1. Xuefei Lu First author E-mail 1471178525@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

2. Xiaoyu Zhang E-mail 774713358@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

3. Zhijie Cao E-mail 1452140040@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

4.Lina Ma **E-mail**✉1902301605@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

5. Jing Ma **E-mail**✉522808421@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

6. Siyi Wang **E-mail**✉1132930923@qq.com

School of Nursing, Xinjiang Medical University, Urumqi 830017, China;

7. Li Zhang ✉Corresponding author✉ **E-mail**✉18099143356@163.com

Department of Nursing, First Affiliated Hospital of Xinjiang Medical University, Urumqi 830054, China;

Health Care Research Center for Xinjiang Regional Population,Urumqi 830017, China;

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors used DeepL Translator solely for initial Chinese-to-English translation of the draft manuscript. No generative AI was used to generate scientific content. All text was fully reviewed and refined by the authors, who take full responsibility for the final submission.

References

1. Singer M, Deutschman C S, Seymour C W, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)[J]. JAMA, 2016,315(8):801-810.DOI:10.1001/jama.2016.0287.
2. Rudd K E, Johnson S C, Agesa K M, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study[J]. Lancet, 2020,395(10219):200-211.DOI:10.1016/S0140-6736(19)32989-7.
3. Gorordo-Delsol L A, Merinos-Sanchez G, Estrada-Escobar R A, et al. Sepsis and septic shock in emergency departments of Mexico: a multicenter point prevalence study[J]. Gac Med Mex, 2020,156(6):486-492.DOI:10.24875/GMM.M21000492.
4. Barichello T, Generoso J S, Singer M, et al. Biomarkers for sepsis: more than just fever and leukocytosis-a narrative review[J]. Crit Care, 2022,26(1):14.DOI:10.1186/s13054-021-03862-5.
5. Povia P, Coelho L, Dal-Pizzol F, et al. How to use biomarkers of infection or sepsis at the bedside: guide to clinicians[J]. Intensive Care Med, 2023,49(2):142-153.DOI:10.1007/s00134-022-06956-y.
6. Novosad S A, Sapiano M R P, Grigg C, et al. Vital Signs: Epidemiology of Sepsis: Prevalence of Health Care Factors and Opportunities for Prevention[J]. MMWR Morb Mortal Wkly Rep, 2016,65(33):864-869.DOI:10.15585/mmwr.mm6533e1.
7. Tang F, Yuan H, Li X, et al. Effect of delayed antibiotic use on mortality outcomes in patients with sepsis or septic shock: A systematic review and meta-analysis[J]. Int Immunopharmacol, 2024,129:111616.DOI:10.1016/j.intimp.2024.111616.
8. Zhou Z, Zhong L, Liu Y, et al. Impact of early heparin therapy on mortality in critically ill patients with sepsis associated acute kidney injury: a retrospective study from the MIMIC-IV database[J]. Front Pharmacol, 2023,14:1261305.DOI:10.3389/fphar.2023.1261305.
9. Arulappen A L, Danial M, Ng L W, et al. The Impact of Antibiotics Administration on Mortality for Time in Sepsis and Septic Shock Patients including Possible Reasons for Delayed Administration in Malaysia[J]. Antibiotics (Basel), 2022,11(9).DOI:10.3390/antibiotics11091202.
10. Seymour C W, Liu V X, Iwashyna T J, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)[J]. JAMA, 2016,315(8):762-774.DOI:10.1001/jama.2016.0288.
11. Chinese College of Emergency Physicians, Shock and Sepsis Professional Committee of Chinese Research Hospital Association. Emergency treatment guidelines for sepsis/septic shock in China (2018). Journal of Clinical Emergency. 2018;19(9):567-588. <https://doi.org/10.13201/j.issn.1009-5918.2018.09.001>
12. Chinese Society of Emergency Medicine, Chinese College of Emergency Physicians, Emergency Medicine Committee of Chinese People's Liberation Army, et al. Expert consensus on emergency fluid therapy for sepsis. Clinical Research and Practice. 2018;3(3):201. <https://doi.org/CNKI:SUN:YLYS.0.2018-03-098>
13. Huang G C, Zhou L, Shi J W, et al. Application and comparison of machine learning algorithms in disease risk prediction. Chinese Health Resources. 2020;23(4):432-436. <https://doi.org/10.13688/j.cnki.chr.2020.19608>

14. Lv X, Liu D, Chen X, et al. Machine learning for the prediction of mortality in patients with sepsis -associated acute kidney injury: a systematic review and meta-analysis[J]. BMC INFECTIOUS DISEASES, 2024,24(1):11.DOI:10.1186/s12879-024-10380-6.
15. Giannini H M, Ginestra J C, Chivers C, et al. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice[J]. Crit Care Med, 2019,47(11):1485-1492.DOI:10.1097/CCM.0000000000003891.
16. Shen E F. Screening of early identification related indicators of sepsis in ICU patients with severe infection[D]. Gannan Medical University, 2024. <https://doi.org/10.27959/d.cnki.ggnyx.2024.000114>
17. Yang D X. Establishment and validation of clinical prediction model for sepsis in ICU patients with suspected infection[D]. Jilin University, 2023. <https://doi.org/10.27162/d.cnki.gjlin.2023.005025>
18. Hu Y, Li M Y, Liu J Q, et al. Construction of a prediction model for short-term ICU readmission in elderly sepsis patients discharged from ICU based on decision tree algorithm. Chinese General Practice Nursing. 2025;23(4):610-615.
19. Wang D D, Zhang L Z, Li X Y. Construction and verification of risk prediction model for sepsis in critically ill ICU patients. Journal of Shanxi Medical University. 2023;54(5):695-701. <https://doi.org/10.13753/j.issn.1007-6611.2023.05.018>
20. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead[J]. Nat Mach Intell, 2019,1(5):206-215.DOI:10.1038/s42256-019-0048-x.
21. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective[J]. BMC Med Inform Decis Mak, 2020,20(1):310.DOI:10.1186/s12911-020-01332-6.
22. Prescott H C, Antonelli M, Alhazzani W, et al. Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock 2026[J]. Intensive Care Medicine, 2026.DOI:10.1007/s00134-026-08361-1.
23. Levy M M, Fink M P, Marshall J C, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference[J]. Crit Care Med, 2003,31(4):1250-1256.DOI:10.1097/01.CCM.0000050454.01978.3B.
24. Infectious Diseases Committee of China Medical Education Association. Expert consensus on clinical significance interpretation of infection-related biomarkers. Chinese Journal of Tuberculosis and Respiratory Diseases. 2017;40(4):243-257. <https://doi.org/10.3760/cma.j.issn.1001-0939.2017.04.002>
25. Procalcitonin Emergency Clinical Application Consensus Group. Expert consensus on emergency clinical application of procalcitonin (PCT). Chinese Journal of Emergency Medicine. 2012;21(9):944-951. <https://doi.org/10.3760/cma.j.issn.1671-0282.2012.09.005>
26. Yealy D M, Mohr N M, Shapiro N I, et al. Early Care of Adults With Suspected Sepsis in the Emergency Department and Out-of-Hospital Environment: A Consensus-Based Task Force Report[J]. Ann Emerg Med, 2021,78(1):1-19.DOI:10.1016/j.annemergmed.2021.02.006.
27. Emergency Medicine Branch of China International Exchange and Promotive Association for Medical and Health Care, Chinese Society of Emergency Medicine, Chinese College of Emergency Physicians, et al. Expert consensus on early prevention and intervention of sepsis in China. 2020.
28. Musat F, Paduraru D N, Bolocan A, et al. Machine Learning Models in Sepsis Outcome Prediction for ICU Patients: Integrating Routine Laboratory Tests-A Systematic Review[J]. Biomedicines, 2024,12(12).DOI:10.3390/biomedicines12122892.
29. Mao Q, Jay M, Hoffman J L, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU[J]. BMJ Open, 2018,8(1):e17833.DOI:10.1136/bmjopen-2017-017833.
30. Riahi A, Yazdani M S, Eshraghi R, et al. Exploring the Potentials of Artificial Intelligence in Sepsis Management in the Intensive Care Unit[J]. Crit Care Res Pract, 2025,2025:9031137.DOI:10.1155/ccrp/9031137.
31. Desai M D, Tootooni M S, Bobay K L. Can Prehospital Data Improve Early Identification of Sepsis in Emergency Department? An Integrative Review of Machine Learning Approaches[J]. Appl Clin Inform, 2022,13(1):189-202.DOI:10.1055/s-0042-1742369.
32. Zhang Y, Xu W, Yang P, et al. Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis[J]. BMC Med Inform Decis Mak, 2023,23(1):283.DOI:10.1186/s12911-023-02383-1.
33. Lauritsen S M, Kalor M E, Kongsgaard E L, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences[J]. Artif Intell Med, 2020,104:101820.DOI:10.1016/j.artmed.2020.101820.

Figures

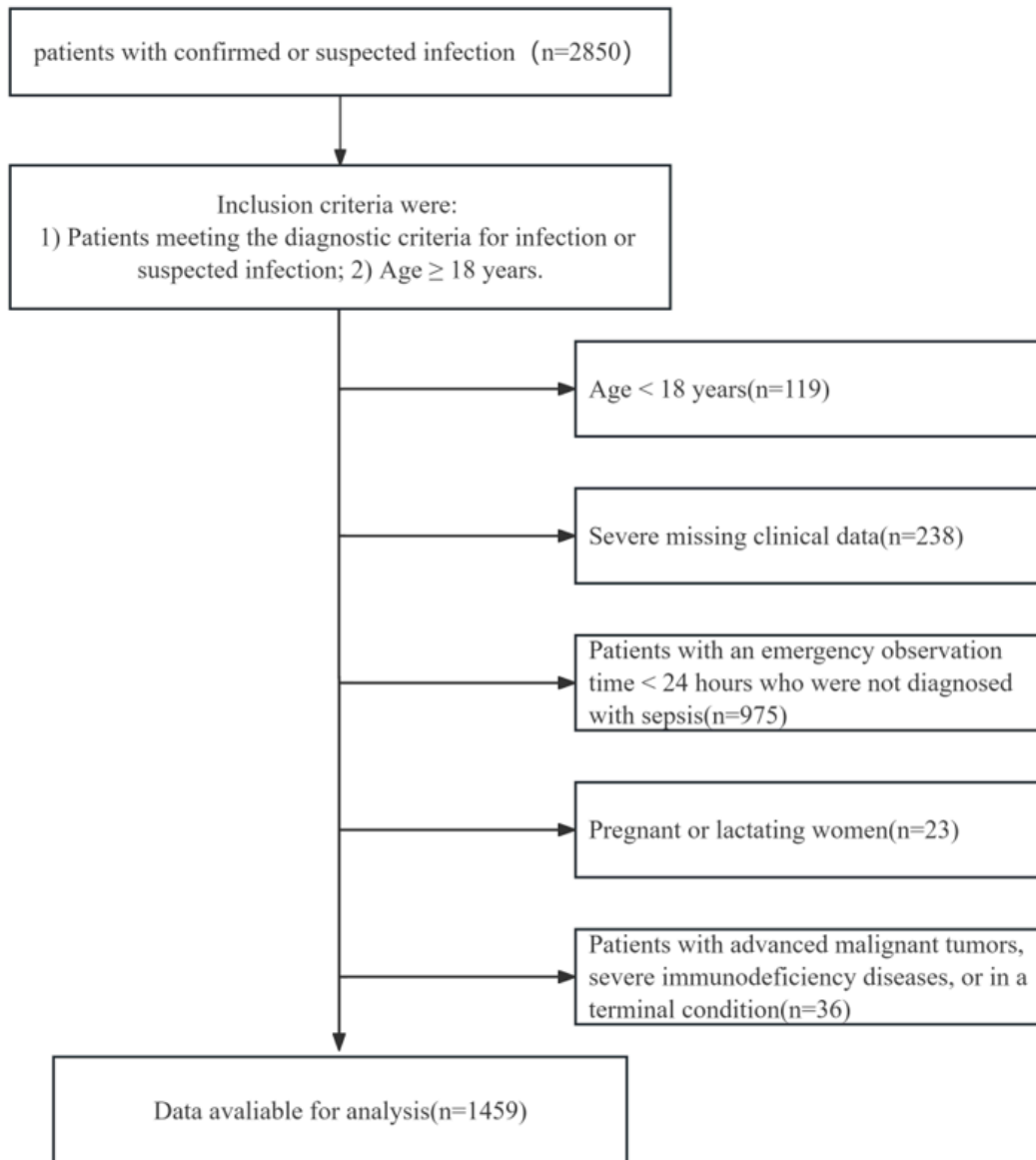


Figure 1

Flowchart of participant selection

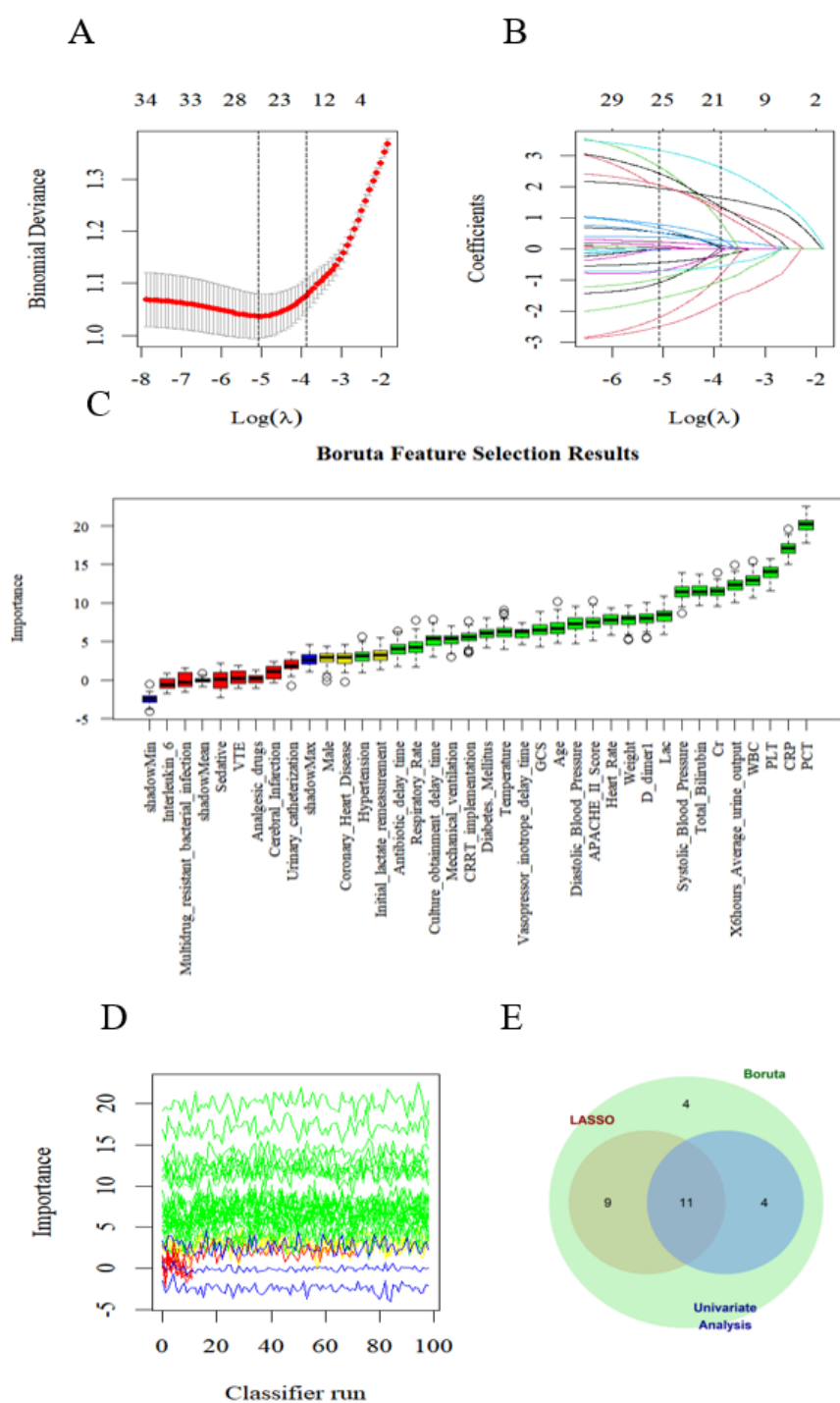


Figure 2

Multi-method feature selection for determining core predictive variables.

Note: (A) LASSO coefficient profiles. (B) Cross-validation curve indicating the optimal lambda (λ_{1se}) for selecting 20 variables. (C) Variable importance ranking by the Boruta algorithm (shadow variables mark thresholds). (D) Stability of feature importance over 100 runs (green: confirmed; yellow: tentative; red: rejected). (E) Venn diagram of variables selected by univariate analysis (15), LASSO (20), and Boruta (28), yielding an 11-variable core set for modeling.

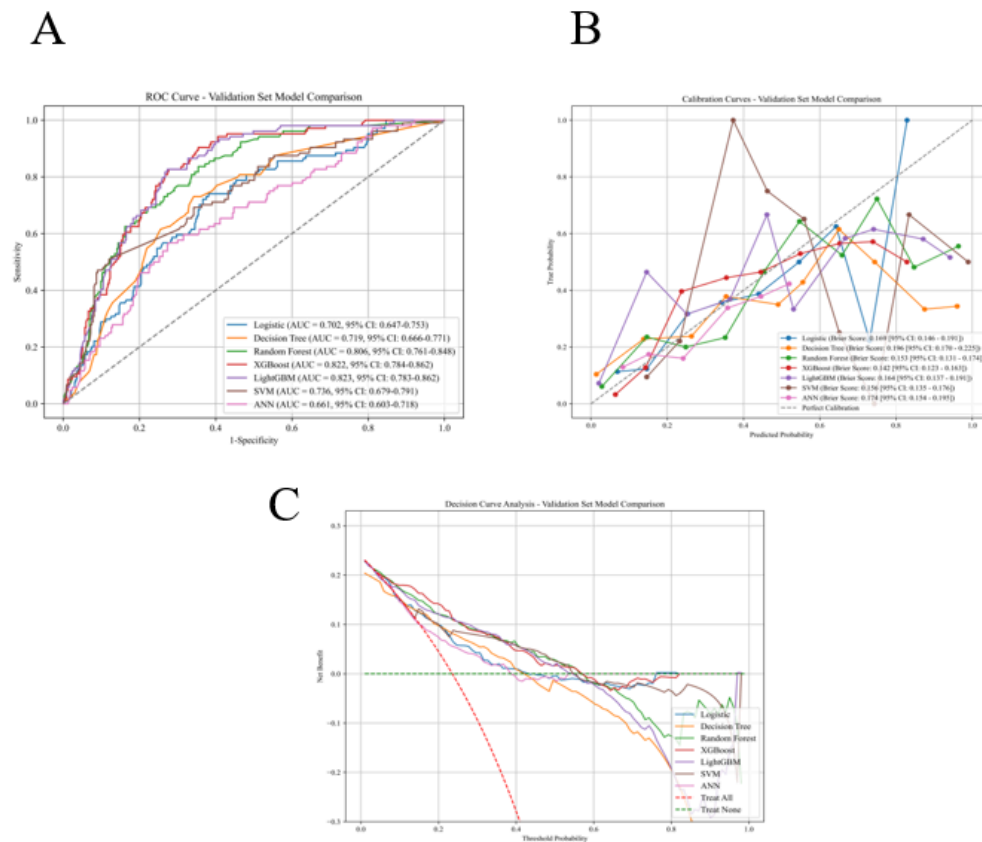


Figure 3

Performance evaluation of machine learning models in the internal validation cohort

Note: (A) Receiver Operating Characteristic (ROC) curves of the seven models. (B) Calibration plots depicting the agreement between predicted probabilities and observed outcomes. (C) Decision Curve Analysis (DCA) evaluating the clinical net benefit across a range of probability thresholds. AUC = Area Under the Curve; Brier = Brier Score.

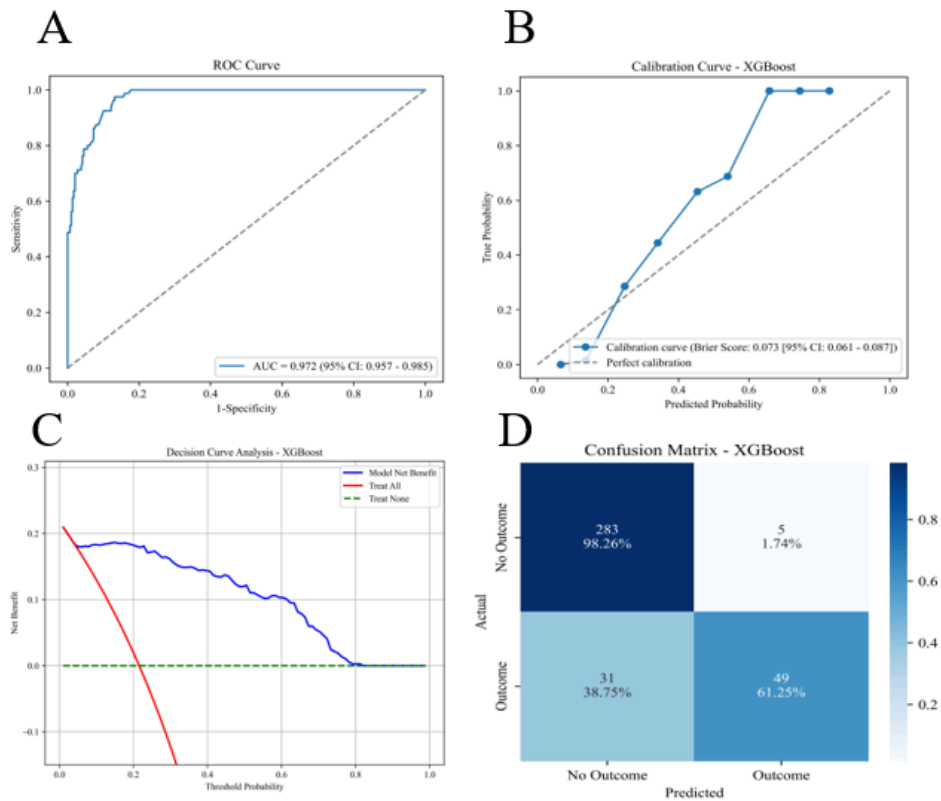


Figure 4

Performance of the XGBoost model in the temporally independent external validation cohort

Note: (A)Receiver operating characteristic (ROC) curve illustrating the model's discrimination capability.(B)Calibration curve showing the alignment between predicted probabilities and observed sepsis incidence.(C)Decision curve analysis (DCA) displaying the net clinical benefit across a range of threshold probabilities.(D)Confusion matrix summarizing classification performance, with the model achieving a specificity of 98.26% and thereby minimizing false positives.

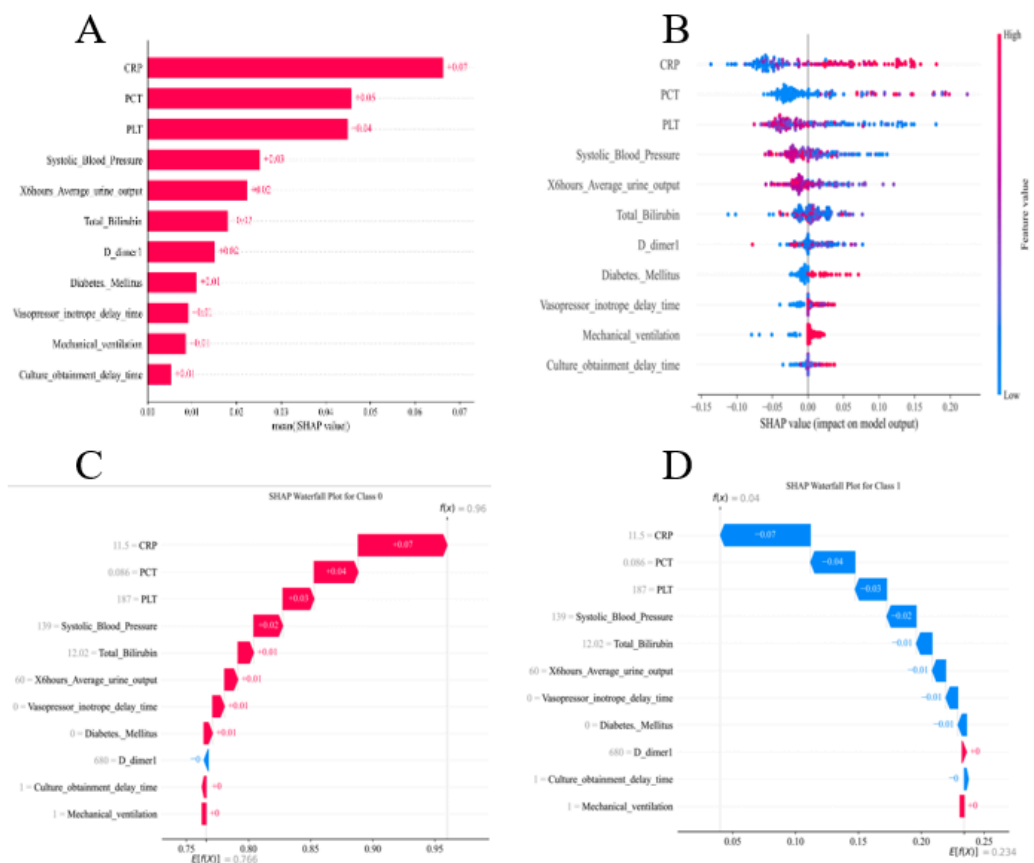


Figure 5

SHAP Explainability Analysis of the XGBoost Sepsis Prediction Model

Note: (A) Global feature importance ranking. (B) SHAP summary plot showing the relationship between feature values (color) and their impact on prediction (horizontal position). (C) Waterfall plot for a representative high-risk sample (Class 0, $f(x)=0.96$). (D) Waterfall plot for a representative low-risk sample (Class 1, $f(x)=0.04$).

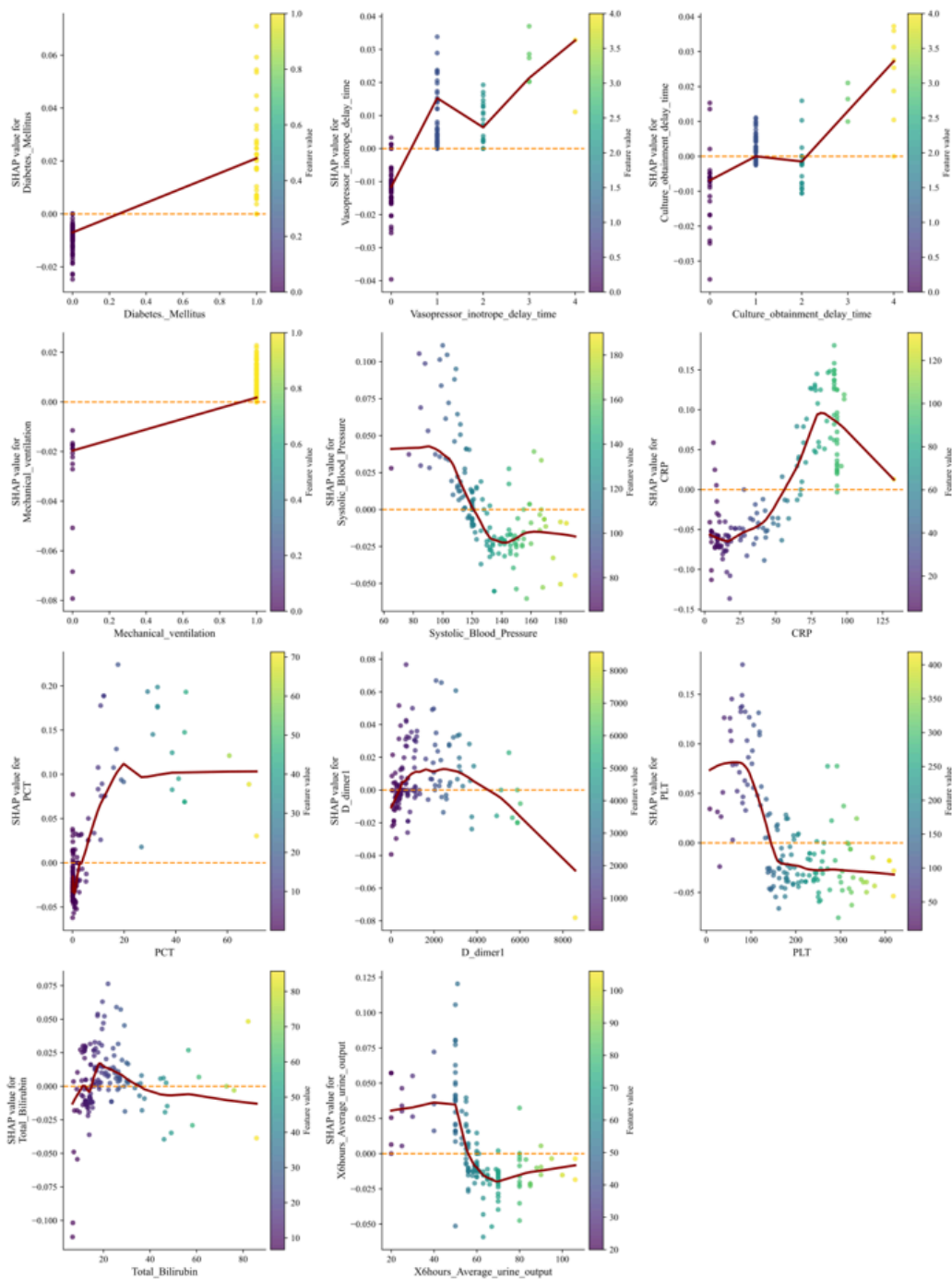


Figure 6

SHAP dependence plots for the 11 core predictors

Note: Features are displayed in descending order of importance, from top to bottom (or panels are arranged left to right, top to bottom): history of diabetes mellitus, time to vasopressor administration, time to blood culture collection, invasive mechanical ventilation, systolic blood pressure, C-reactive protein (CRP), procalcitonin (PCT), D-dimer, platelet count, total bilirubin, and 6-hour average urine output.