

An interpretable machine learning model for early sepsis prediction in patients with heart failure and type 2 diabetes: a retrospective cohort study

Hanchen Wu

Second Military Medical University

Yifan Wang

Second Military Medical University

Hanbai Li

Second Military Medical University

Xin Xie

Second Military Medical University

Xinrui Shang

Second Military Medical University

Zhengyu Jiang

jzysmmu@outlook.com

The First Affiliated Hospital of Naval Medical University

Zhanheng Chen

Second Military Medical University

Research Article

Keywords: heart failure, type 2 diabetes mellitus, sepsis, machine learning, XGBoost

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10701420/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

Background

Patients with both heart failure (HF) and type 2 diabetes mellitus (T2DM) are susceptible to sepsis, but no validated tool addresses early risk stratification in this population. We developed and validated interpretable machine learning models using two critical care databases.

Methods

Adult intensive care unit patients with HF and T2DM were identified in MIMIC-IV (2008–2022); an eICU cohort was used for external validation. Feature selection combined univariate and multivariable analyses, least absolute shrinkage and selection operator regression, and the Boruta algorithm. Seven models were tuned by grid search with five-fold cross-validation. Discrimination, calibration, clinical utility, and interpretability were assessed with the area under the receiver operating characteristic curve (AUC-ROC), Brier score, decision curve analysis, and SHapley Additive exPlanations (SHAP), respectively.

Results

The MIMIC-IV cohort comprised 7,806 patients, of whom 4,195 (53.74%) had sepsis; the external eICU cohort comprised 633 patients. Seven predictors were retained: acute kidney injury, mechanical ventilation, white blood cell count, body temperature, Glasgow Coma Scale score, anion gap, and blood lactate. XGBoost had the highest overall performance, with AUC-ROC values of 0.712 (95% CI: 0.691–0.733) in internal validation and 0.658 (95% CI: 0.551–0.756) in external validation. SHAP ranked acute kidney injury, Glasgow Coma Scale score, white blood cell count, mechanical ventilation, and body temperature as the leading contributors.

Conclusions

The XGBoost model provided moderate discrimination with interpretable predictions based on routinely available clinical variables. It may support early sepsis risk assessment in intensive care patients with HF and T2DM, but prospective multicenter validation is needed before clinical use.

Background

Heart failure (HF) often follows a prolonged course with recurrent acute decompensation. Coexisting type 2 diabetes mellitus (T2DM) adds cardiorenal and metabolic dysfunction and is associated with greater risks of infection and sepsis [1]. Patients with diabetes are more susceptible to pulmonary, urinary tract, and bloodstream infections [2]. Sepsis also carries high rates of organ failure, intensive care unit (ICU) admission, mortality, and healthcare use [3, 4]. In patients with HF and T2DM, chronic inflammation, impaired immunity, cardiorenal dysfunction, altered hemodynamics, poor microcirculatory perfusion, and hyperglycemia-related endothelial injury may contribute to this risk [5]. Advanced age, multimorbidity, and polypharmacy can further obscure early clinical deterioration.

Sepsis risk assessment commonly relies on general scores such as the Systemic Inflammatory Response Syndrome criteria, quick Sequential Organ Failure Assessment, Sequential Organ Failure Assessment, and Acute Physiology and Chronic Health Evaluation II, together with individual inflammatory or infection markers [6]. These tools may not capture nonlinear associations among organ dysfunction, metabolic abnormalities, treatment exposures, and infection-related variables. No prediction model has been developed specifically for ICU patients with both HF and T2DM.

Machine learning can model nonlinear patterns in high-dimensional electronic health record data [7]. Previous models have predicted sepsis [8, 9] and readmission or mortality in HF [10, 11], while other studies have examined cardiovascular risk in patients with diabetes [12, 13]. For clinical use, performance must be paired with interpretable estimates of how each predictor contributes to an individual prediction. SHapley Additive exPlanations (SHAP) provide one such approach [14].

Reliable development also requires transparent methods and independent validation across centers [15, 16]. We therefore developed and externally validated seven machine learning models for sepsis risk in ICU patients with HF and T2DM. We compared discrimination, calibration, and clinical utility, then interpreted the best-performing model with SHAP.

Methods

Study design and data sources

This retrospective cohort study used two publicly available critical care databases: Medical Information Mart for Intensive Care IV (MIMIC-IV, version 3.1) and the eICU Collaborative Research Database. MIMIC-IV contains admissions to Beth Israel Deaconess Medical Center in the United States from 2008 to 2022. The eICU database includes patients treated in 335 ICUs at 208 US hospitals during 2014–2015 and was used for external validation.

Patients were eligible if they were aged at least 18 years, had confirmed HF and T2DM, were undergoing their first ICU admission, and remained in the ICU for at least 12 hours. Patients with end-stage renal disease (chronic kidney disease stage 5), human immunodeficiency virus infection, or brain death were excluded. The study used de-identified public data and was conducted in accordance with the 2013 revision of the Declaration of Helsinki. The cohort construction process is summarized in Fig. 1.

[Insert Fig. near here]

Outcome definition

HF and T2DM were identified with International Classification of Diseases, 10th Revision codes I50 and E11, respectively. Sepsis was defined according to Sepsis-3 as documented or suspected infection with an acute increase of at least 2 points in the total Sequential Organ Failure Assessment score [17]. Infection was identified from ICD codes.

Candidate variables

Variables recorded during the first 24 hours after ICU admission were extracted from MIMIC-IV. Baseline variables were age, sex, body weight, hypertension, malignancy, cerebrovascular accident, acute kidney injury (AKI), liver cirrhosis, mechanical ventilation (MV), and ICU length of stay. Vital signs comprised heart rate, systolic and diastolic blood pressure, mean arterial pressure, respiratory rate, peripheral oxygen saturation, and body temperature. Severity measures were the Sequential Organ Failure Assessment score, Acute Physiology and Chronic Health Evaluation III score, Systemic Inflammatory Response Syndrome score, Simplified Acute Physiology Score II, Oxford Acute Severity of Illness Score, and Glasgow Coma Scale (GCS) score. Laboratory variables included hematocrit, hemoglobin, platelet count, red cell distribution width, red blood cell count, white blood cell (WBC) count, anion gap (AG), total calcium, chloride, glucose, potassium, sodium, bicarbonate, magnesium, phosphate, lactate, partial pressures of carbon dioxide and oxygen, blood pH, international normalized ratio, prothrombin time, activated partial thromboplastin time, creatinine, and blood urea nitrogen.

The same variables were extracted from eICU where available. Sequential Organ Failure Assessment, Acute Physiology and Chronic Health Evaluation III, Systemic Inflammatory Response Syndrome, Simplified Acute Physiology Score II, and Oxford Acute Severity of Illness scores were not extracted from eICU.

Data preprocessing

MIMIC-IV variables with more than 30% missing observations were excluded. Values more than 3 standard deviations from the mean were treated as missing, and remaining missing values were multiply imputed [18].

For external validation, patients missing any model predictor were excluded. Values more than 3 standard deviations from the mean were treated as missing, and the remaining missing values were multiply imputed.

Feature selection

The MIMIC-IV cohort was randomly divided 7:3 into training and test sets. Three feature-selection methods were applied independently in the training set.

First, variables with $P < 0.05$ in univariate analyses entered a multivariable logistic regression model. Second, least absolute shrinkage and selection operator regression used an L1 penalty, with 10-fold cross-validation to select λ ; variables with nonzero coefficients were retained [19]. Third, the Boruta algorithm compared the random-forest importance Z-score of each observed feature with those of randomized shadow features and removed features that did not exceed random noise [20].

The intersection of the three feature sets defined the core candidate set. This approach combined hypothesis-based screening, penalized regression, and random-forest importance.

Model development

Seven supervised algorithms were evaluated: logistic regression, random forest, support vector machine, decision tree, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and artificial neural network. Models were implemented in Python 3.9 or later with scikit-learn 1.2 or later, lightgbm, and xgboost. Hyperparameters were selected by grid search with five-fold cross-validation in the training set.

Model evaluation

Discrimination was assessed with receiver operating characteristic curves and the area under the curve (AUC), supplemented by accuracy, sensitivity, specificity, precision, F1 score, Cohen kappa, Youden index, positive predictive value, and negative predictive value. Calibration was examined with calibration curves and the Brier score, for which lower values indicate smaller prediction error. Decision curve analysis quantified net benefit across threshold probabilities [21].

Model interpretation

The best-performing model was interpreted with SHAP values [22]. Mean absolute SHAP values ranked global feature importance, and a beeswarm plot showed the direction and distribution of feature contributions. Dependence plots examined nonlinear associations between predictor values and model output. Waterfall and force plots illustrated how features contributed to predictions for representative patients.

Sample size

Following Riley et al. [23], the sample-size assessment targeted a mean absolute prediction error of 0.05. A cohort of 7,806 patients was considered sufficient for the 9 predictor parameters specified in the calculation.

Statistical analysis

Normally distributed continuous variables are reported as mean $\pm$ standard deviation, nonnormally distributed variables as median (interquartile range), and categorical variables as number (percentage). Continuous variables were compared with the independent-samples t test or Kruskal-Wallis test, as appropriate. Categorical variables were compared with the chi-square test or Fisher exact test. Tests were two-sided, and $P < 0.05$ was considered statistically significant. Analyses were performed in R 4.2.1 (R Foundation for Statistical Computing). The glmnet and Boruta packages were used for feature selection, caret for model training and grid search, and shapviz and lime for model interpretation.

Use of AI-assisted technologies

OpenAI ChatGPT was used to support English-language editing and manuscript formatting. It was not used to generate or analyze study data. The authors reviewed all changes and take responsibility for the final content.

Results

Patient characteristics

MIMIC-IV contained 9,060 potentially eligible patients with HF and T2DM. After 1,254 patients with end-stage renal disease, human immunodeficiency virus infection, or brain death were excluded, 7,806 remained; 4,195 (53.74%) met the sepsis definition.

The training set contained 5,465 patients, including 2,937 (53.74%) with sepsis. In the sepsis group, mean age was 73.60 ± 11.02 years, 56.69% were male, 65.27% had AKI, and 88.93% received MV. Median WBC count was $11.40 (8.20, 15.80) \times 10^9/L$, arterial PO_2 was 81.00 (48.00, 157.00) mmHg, blood urea nitrogen was 34.00 (21.00, 52.00) mmol/L, and mean magnesium was 2.04 ± 0.42 mmol/L. Age, body weight, peripheral oxygen saturation, PCO_2 , platelet count, vasoactive drug use, glucocorticoid use, antibiotic use, cerebrovascular disease, and hypertension did not differ between the sepsis and nonsepsis groups (all $P > 0.05$); the remaining baseline variables differed (all $P < 0.05$) (Table 1).

Table 1 Baseline characteristics of the training cohort.

Variables	Total (n = 5465)	Non-sepsis (n = 2528)	Sepsis (n = 2937)	Statistic	P
Age, years	73.46 ± 11.21	73.30 ± 11.43	73.60 ± 11.02	t=-1.00	0.317
Weight, kg	88.69 ± 23.81	88.84 ± 24.22	88.55 ± 23.45	t=0.44	0.658
GCS, score	14.06 ± 1.80	14.32 ± 1.49	13.84 ± 2.00	t=10.26	<0.001
HR, beats/min	86.56 ± 19.05	84.60 ± 17.87	88.25 ± 19.85	t=-7.14	<0.001
SBP, mmHg	121.91 ± 23.99	124.03 ± 24.07	120.08 ± 23.78	t=6.09	<0.001
DBP, mmHg	66.38 ± 16.97	67.61 ± 16.77	65.32 ± 17.08	t=4.97	<0.001
MAP, mmHg	80.66 ± 17.44	82.15 ± 17.10	79.38 ± 17.63	t=5.88	<0.001
RR, breaths/min	19.54 ± 5.63	19.11 ± 5.45	19.91 ± 5.75	t=-5.33	<0.001
SPO2, %	96.84 ± 3.29	96.93 ± 3.12	96.77 ± 3.42	t=1.78	0.076
Temperature, °C	36.74 ± 0.66	36.67 ± 0.53	36.80 ± 0.74	t=-7.84	<0.001
Hematocrit, %	31.47 ± 6.39	31.67 ± 6.53	31.30 ± 6.27	t=2.17	0.030
Hemoglobin, g/dL	10.10 ± 2.12	10.19 ± 2.16	10.02 ± 2.08	t=2.95	0.003
RDW, %	15.77 ± 2.19	15.61 ± 2.17	15.91 ± 2.20	t=-5.08	<0.001
RBC, ×10 ¹² /L	3.51 ± 0.76	3.55 ± 0.77	3.48 ± 0.75	t=3.21	0.001
AG, mmol/L	14.74 ± 3.89	14.23 ± 3.55	15.18 ± 4.10	t=-9.21	<0.001
First Calcium Total, mg/dL	8.50 ± 0.67	8.63 ± 0.62	8.39 ± 0.69	t=13.16	<0.001
Chloride, mmol/L	102.14 ± 6.49	101.95 ± 6.10	102.31 ± 6.80	t=-2.04	0.042
Potassium, mmol/L	4.34 ± 0.71	4.31 ± 0.66	4.37 ± 0.76	t=-3.24	0.001
Sodium, mmol/L	138.09 ± 4.82	138.06 ± 4.44	138.11 ± 5.13	t=-0.42	0.677
Bicarbonate, mmol/L	23.75 ± 5.39	24.36 ± 5.00	23.23 ± 5.65	t=7.86	<0.001
Magnesium, mg/dL	2.05 ± 0.41	2.07 ± 0.40	2.04 ± 0.42	t=2.98	0.003
Phosphate, mg/dL	3.94 ± 1.18	3.86 ± 1.04	4.01 ± 1.28	t=-4.75	<0.001
PCO ₂ , mmHg	44.97 ± 12.38	44.91 ± 12.01	45.03 ± 12.68	t=-0.36	0.717
pH	7.37 ± 0.09	7.38 ± 0.08	7.36 ± 0.09	t=11.08	<0.001
Platelet, ×10 ⁹ /L	197.00 (147.00, 258.00)	199.00 (152.00, 255.00)	195.00 (141.00, 261.00)	Z=-1.90	0.058
WBC, ×10 ⁹ /L	10.40 (7.60, 14.40)	9.50 (7.20, 12.60)	11.40 (8.20, 15.80)	Z=-13.59	<0.001
Glucose, mmol/L	8.44 (6.50, 11.44)	8.11 (6.39, 10.94)	8.72 (6.61, 12.00)	Z=-5.45	<0.001
Lactate, mmol/L	1.60 (1.10, 2.40)	1.50 (1.10, 2.20)	1.70 (1.20, 2.60)	Z=-7.37	<0.001
PO ₂ , mmHg	82.00 (48.00, 166.00)	84.00 (48.00, 179.25)	81.00 (48.00, 157.00)	Z=-2.04	0.042
INR	1.30 (1.20, 1.70)	1.30 (1.20, 1.60)	1.40 (1.20, 1.80)	Z=-6.64	<0.001
PT, s	14.80 (12.90, 18.20)	14.40 (12.70, 17.40)	15.10 (13.10, 19.10)	Z=-7.04	<0.001
APTT, s	31.70 (27.80, 38.60)	31.30 (27.70, 37.90)	32.00 (27.90, 39.10)	Z=-2.28	0.023
Creatinine, mg/dL	1.30 (1.00, 2.00)	1.20 (0.90, 1.80)	1.50 (1.00, 2.20)	Z=-10.73	<0.001
First Urea Nitrogen, mg/dL	30.00 (19.00, 48.00)	27.00 (17.00, 43.00)	34.00 (21.00, 52.00)	Z=-10.70	<0.001
Vp, n(%)				χ ² =0.29	0.591
No	2862 (52.37)	1314 (51.98)	1548 (52.71)		
Yes	2603 (47.63)	1214 (48.02)	1389 (47.29)		
Gc, n(%)				χ ² =0.29	0.588
No	4070 (74.47)	1874 (74.13)	2196 (74.77)		
Yes	1395 (25.53)	654 (25.87)	741 (25.23)		
Abx, n(%)				χ ² =0.45	0.504

Variables	Total (n = 5465)	Non-sepsis (n = 2528)	Sepsis (n = 2937)	Statistic	P
No	1148 (21.01)	521 (20.61)	627 (21.35)		
Yes	4317 (78.99)	2007 (79.39)	2310 (78.65)		
AKI, n(%)				$\chi^2=196.08$	<0.001
No	2374 (43.44)	1354 (53.56)	1020 (34.73)		
Yes	3091 (56.56)	1174 (46.44)	1917 (65.27)		
Lc, n(%)				$\chi^2=4.78$	0.029
No	5196 (95.08)	2421 (95.77)	2775 (94.48)		
Yes	269 (4.92)	107 (4.23)	162 (5.52)		
CVA, n(%)				$\chi^2=0.01$	0.907
No	4731 (86.57)	2187 (86.51)	2544 (86.62)		
Yes	734 (13.43)	341 (13.49)	393 (13.38)		
Tumor, n(%)				$\chi^2=4.35$	0.037
No	4613 (84.41)	2106 (83.31)	2507 (85.36)		
Yes	852 (15.59)	422 (16.69)	430 (14.64)		
Hypertension, n(%)				$\chi^2=0.14$	0.705
No	4129 (75.55)	1904 (75.32)	2225 (75.76)		
Yes	1336 (24.45)	624 (24.68)	712 (24.24)		
Male, n(%)				$\chi^2=0.32$	0.574
No	2386 (43.66)	1114 (44.07)	1272 (43.31)		
Yes	3079 (56.34)	1414 (55.93)	1665 (56.69)		
MV, n(%)				$\chi^2=158.42$	<0.001
No	929 (17.00)	604 (23.89)	325 (11.07)		
Yes	4536 (83.00)	1924 (76.11)	2612 (88.93)		

Abbreviations: GCS, Glasgow Coma Scale; HR, heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; RR, respiratory rate; SPO₂, peripheral oxygen saturation; RDW, red cell distribution width; RBC, red blood cell count; AG, anion gap; PCO₂, partial pressure of carbon dioxide; INR, international normalized ratio; PT, prothrombin time; APTT, activated partial thromboplastin time; AKI, acute kidney injury; MV, mechanical ventilation; Vp, vasoactive drugs; Gc, glucocorticoids; Abx, antibiotics; Lc, liver cirrhosis; CVA, cerebrovascular accident.

Feature selection

Univariate screening followed by multivariable logistic regression retained 20 variables associated with sepsis in the training set (P<0.05) (Table 2).

Table 2 Univariate screening and multivariable logistic regression of sepsis predictors.

Variable	Beta	SE	Wald	OR	95% CI lower	95% CI upper	P
GCS	-0.233	0.025	-9.182	0.792	0.753	0.832	<0.001
HR	0.005	0.002	2.707	1.005	1.001	1.008	0.007
SBP	-0.001	0.002	-0.384	0.999	0.995	1.003	0.701
DBP	0.001	0.005	0.194	1.001	0.991	1.011	0.846
MAP	-0.006	0.006	-1.02	0.994	0.982	1.006	0.308
RR	0.004	0.006	0.627	1.004	0.992	1.016	0.53
Temperature	0.354	0.058	6.066	1.424	1.271	1.597	<0.001
Hematocrit	0.005	0.022	0.231	1.005	0.963	1.049	0.817
Hemoglobin	0.088	0.058	1.511	1.092	0.974	1.223	0.131
RDW	0.056	0.017	3.228	1.058	1.022	1.095	0.001
RBC	-0.3	0.106	-2.823	0.74	0.601	0.912	0.005
AG	0.035	0.011	3.081	1.035	1.013	1.059	0.002
Total calcium	-0.506	0.053	-9.484	0.603	0.543	0.669	<0.001
Chloride	0.018	0.006	2.837	1.018	1.005	1.03	0.005
Potassium	0.006	0.05	0.129	1.007	0.912	1.111	0.897
Bicarbonate	0.025	0.009	2.862	1.025	1.008	1.043	0.004
Magnesium	-0.474	0.085	-5.576	0.623	0.527	0.735	<0.001
Phosphate	-0.086	0.034	-2.499	0.918	0.858	0.982	0.012
pH	-1.809	0.42	-4.306	0.164	0.072	0.373	<0.001
WBC	0.056	0.006	8.699	1.058	1.044	1.071	<0.001
Glucose	0.027	0.009	3.094	1.027	1.01	1.045	0.002
Lactate	0.079	0.032	2.443	1.082	1.016	1.153	0.015
PO ₂	0	0	0.544	1	1	1.001	0.586
INR	0.005	0.114	0.04	1.005	0.803	1.257	0.968
PT	0.017	0.011	1.567	1.018	0.996	1.04	0.117
Creatinine	0.142	0.062	2.286	1.152	1.02	1.301	0.022
Blood urea nitrogen	0.007	0.002	3.171	1.007	1.003	1.012	0.002
AKI	0.548	0.069	7.989	1.73	1.513	1.98	<0.001
Liver cirrhosis	0.17	0.143	1.187	1.185	0.895	1.568	0.235
MV	0.955	0.084	11.304	2.599	2.202	3.067	<0.001
APTT	0.003	0.003	1.124	1.003	0.998	1.009	0.261
Malignancy	-0.182	0.082	-2.212	0.834	0.71	0.98	0.027

Boruta retained 30 potential predictors, and least absolute shrinkage and selection operator regression retained 18 variables at the 1-standard-error criterion. Fifteen variables appeared in the intersection of all three methods. From this set, seven clinically representative features were selected for model development: AKI, MV, WBC count, body temperature, GCS score, AG, and blood lactate. The complete feature-selection results are shown in Fig. 2.

Model performance

Seven models were evaluated in the training, test, and external validation sets (Tables 3-5).

Table 3 Model performance in the training set.

Model	AUC	AUC 95% CI Lower	AUC 95% CI Upper	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Kappa	Youden's J	PPV	NPV
Logistic	0.7025	0.6878	0.7172	0.6456	0.6581	0.7085	0.5724	0.6824	0.2827	0.2809	0.6581	0.6283
Decision Tree	0.7144	0.7007	0.7275	0.6586	0.6583	0.7583	0.5427	0.7047	0.3047	0.3010	0.6583	0.6590
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
XGBoost	0.7447	0.7309	0.7574	0.6783	0.6899	0.7293	0.6191	0.7090	0.3500	0.3484	0.6899	0.6631
LightGBM	0.8704	0.8613	0.8792	0.7757	0.7782	0.8148	0.7302	0.7961	0.5471	0.5450	0.7782	0.7724
SVM	0.7003	0.6856	0.7147	0.6467	0.6574	0.7154	0.5669	0.6851	0.2842	0.2822	0.6574	0.6316
ANN	0.7192	0.7050	0.7320	0.6595	0.6740	0.7096	0.6013	0.6913	0.3122	0.3108	0.6740	0.6405

Table 4 Model performance in the test set.

Model	AUC	AUC 95% CI Lower	AUC 95% CI Upper	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Kappa	Youden's J	PPV	NPV
Logistic	0.6979	0.6782	0.7197	0.6467	0.6532	0.7305	0.5494	0.6897	0.2826	0.2799	0.6532	0.6370
Decision Tree	0.6765	0.6560	0.6976	0.6322	0.6368	0.7345	0.5134	0.6822	0.2510	0.2479	0.6368	0.6247
Random Forest	0.6758	0.6540	0.6973	0.6292	0.6419	0.7011	0.5457	0.6702	0.2487	0.2468	0.6419	0.6112
XGBoost	0.7117	0.6907	0.7328	0.6578	0.6722	0.7091	0.5983	0.6901	0.3088	0.3074	0.6722	0.6391
LightGBM	0.6977	0.6770	0.7192	0.6420	0.6556	0.7035	0.5706	0.6787	0.2758	0.2741	0.6556	0.6236
SVM	0.6968	0.6772	0.7185	0.6484	0.6546	0.7321	0.5512	0.6912	0.2861	0.2834	0.6546	0.6392
ANN	0.6994	0.6785	0.7218	0.6514	0.6635	0.7130	0.5799	0.6874	0.2947	0.2929	0.6635	0.6350

Table 5 Model performance in the external validation set.

Model	AUC	AUC 95% CI Lower	AUC 95% CI Upper	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Kappa	Youden's J	PPV	NPV
Logistic	0.6713	0.5543	0.7753	0.5229	0.9781	0.5140	0.7308	0.6739	0.0388	0.2448	0.9781	0.0605
Decision Tree	0.6065	0.4529	0.7388	0.6003	0.9733	0.5997	0.6154	0.7421	0.0407	0.2151	0.9733	0.0618
Random Forest	0.6213	0.5228	0.7161	0.6130	0.9665	0.6178	0.5000	0.7538	0.0234	0.1178	0.9665	0.0531
XGBoost	0.6578	0.5506	0.7561	0.6161	0.9715	0.6178	0.5769	0.7553	0.0384	0.1947	0.9715	0.0607
LightGBM	0.6101	0.5047	0.7113	0.5893	0.9677	0.5914	0.5385	0.7342	0.0243	0.1299	0.9677	0.0534
SVM	0.6989	0.5889	0.7970	0.5419	0.9789	0.5338	0.7308	0.6908	0.0435	0.2645	0.9789	0.0629
ANN	0.6895	0.5829	0.7818	0.5940	0.9781	0.5898	0.6923	0.7359	0.0519	0.2821	0.9781	0.0674

Comparison of receiver operating characteristic curves, calibration curves, and decision curves identified XGBoost as the best-performing model (Fig. 3).

In the test set, XGBoost yielded an AUC of 0.712 (95% CI: 0.691-0.733), accuracy of 65.78%, sensitivity of 70.91%, precision of 67.22%, specificity of 59.83%, and an F1 score of 0.690. In external validation, the AUC was 0.6578 (95% CI: 0.5506-0.7561), accuracy was 61.61%, sensitivity was 61.78%, positive predictive value was 97.15%, specificity was 57.69%, and the F1 score was 0.7553.

Model interpretation

Mean absolute SHAP values ranked AKI, GCS score, and WBC count as the three largest contributors to XGBoost predictions, followed by MV, body temperature, AG, and lactate (Fig. 4A).

The beeswarm plot showed the direction and magnitude of each feature contribution (Fig. 4B). Higher WBC counts, AKI, and MV generally shifted predictions toward sepsis, whereas higher GCS scores shifted them away from sepsis. Body temperature, AG, and lactate showed nonlinear associations with model output (Fig. 6).

Force plots illustrated two individual predictions (Fig. 5). For the high-risk example, the baseline probability was 0.3872; MV=1, lactate=4.3, and AKI=1 increased the predicted probability to 0.5372. For the lower-risk example, the baseline probability was 0.3128. Despite AKI=1, lactate=4.3, and MV=1, WBC=11.9 and temperature=36.67 did not add substantial risk, and the final predicted probability was 0.342.

Discussion

XGBoost provided the best overall performance among the seven models. Its AUC was 0.712 (95% CI: 0.691-0.733) in the test set and 0.658 (95% CI: 0.5506-0.7561) in external validation. The model used seven routinely recorded variables and provided both global and patient-level explanations. These results support XGBoost as a candidate risk-stratification method, although its moderate external discrimination does not yet support autonomous clinical decisions.

Feature selection combined three statistical and machine learning approaches, followed by clinical review. SHAP ranked AKI, GCS score, and WBC count highest. Dependence plots suggested positive contributions from AKI and MV, a U-shaped association with body temperature, a plateau in the contribution of WBC above approximately $15 \times 10^9/L$, and a stronger positive contribution from lactate above approximately 3 mmol/L. Higher GCS scores contributed to lower predicted risk. AG showed a negative contribution below approximately 15 mmol/L.

AKI was a leading model feature and increased predicted sepsis risk. AKI is common in sepsis and is associated with organ failure and mortality [24]. Inflammation, disrupted microcirculatory perfusion, and endothelial injury contribute to sepsis-associated AKI [25]. These processes can injure renal tubular cells and redistribute renal blood flow, while impaired renal function may intensify systemic inflammation and multiple-organ dysfunction [26]. In patients with HF and T2DM, AKI may therefore mark both organ injury and more severe infection-related illness.

MV also contributed positively to sepsis predictions. It often marks severe respiratory failure or greater overall illness severity and is associated with infection risk in ICU patients [27]. Endotracheal intubation disrupts airway defenses and facilitates lower respiratory tract colonization, which increases the risk of ventilator-associated pneumonia [28]. Ventilation may also promote pulmonary and systemic inflammatory responses [29]. The association should nevertheless be interpreted as predictive rather than causal.

WBC count remained informative despite its limited specificity. Infection activates innate immune signaling and stimulates leukocyte release from bone marrow [30]. WBC count is therefore included in several infection-related scores and diagnostic criteria [31]. Persistent leukocyte abnormalities may also reflect dysregulated immunity during severe infection and sepsis [32].

Body temperature showed a U-shaped association with model output. Fever reflects cytokine effects on hypothalamic thermoregulation [33], whereas hypothermia in critically ill patients is often associated with more severe illness and poorer outcomes [34]. Both departures from the normal range may therefore aid early sepsis risk assessment [35].

Higher lactate levels were associated with higher predicted risk. Lactate reflects tissue perfusion and cellular metabolism [36]. During severe infection or circulatory dysfunction, insufficient oxygen delivery can increase anaerobic metabolism and lactate production [37]. Elevated lactate is also associated with sepsis severity and mortality [38].

Higher GCS scores contributed to lower predicted risk. Sepsis may alter consciousness through systemic inflammation, cerebral hypoperfusion, and blood-brain barrier injury, producing sepsis-associated encephalopathy [39]. Altered mental status is common in severe infection and is associated with mortality [40]. GCS therefore captures neurologic function as well as illness severity [41].

AG contributed nonlinearly to predictions. Metabolic acidosis in sepsis may reflect tissue hypoxia, lactate accumulation, and other metabolic disturbances [42]. Accumulated organic acids raise the AG [43], so AG may indirectly capture impaired tissue perfusion and illness severity [44].

This study focused on a population that is not represented by a dedicated sepsis risk model. XGBoost can represent nonlinear associations among multiple clinical variables, and SHAP made those associations visible at both cohort and patient levels. The seven predictors are routinely available in ICU records, which may ease implementation if the model is validated prospectively.

The cohort size supported model development, but performance in external validation was moderate rather than high. The online implementation should therefore be viewed as a research tool for risk estimation, not as a basis for treatment decisions. A cloud-based interface could help clinicians identify patients who warrant closer assessment, but any intervention must remain guided by clinical evaluation.

Several limitations affect interpretation. First, the retrospective design may have introduced selection bias. Second, MIMIC-IV represents one US center, and the external cohort may not capture other health systems or populations. Third, admissions spanned 2008-2022, during which treatment practices

changed; the cohort was also nonconsecutive. Fourth, only seven commonly used algorithms were evaluated, so another method may perform better. Prospective multicenter studies should examine temporal validity, calibration, and clinical effects before deployment.

Conclusions

Among seven machine learning models, XGBoost provided the best discrimination for sepsis risk in ICU patients with HF and T2DM. The model used seven clinical variables and produced interpretable patient-level predictions. Its moderate external performance warrants prospective multicenter validation before clinical use.

Abbreviations

AG, Anion Gap; AKI, Acute Kidney Injury; ANN, Artificial Neural Network; APACHE II, Acute Physiology and Chronic Health Evaluation II; APTT, Activated Partial Thromboplastin Time; AUC, Area Under the Curve; AUC-ROC, Area Under the Receiver Operating Characteristic Curve; CI, Confidence Interval; CKD, Chronic Kidney Disease; CVA, Cerebrovascular Accident; DBP, Diastolic Blood Pressure; DCA, Decision Curve Analysis; DT, Decision Tree; GCS, Glasgow Coma Scale; HF, Heart Failure; HIV, Human Immunodeficiency Virus; HR, Heart Rate; ICD-10, International Classification of Diseases, 10th Revision; ICU, Intensive Care Unit; INR, International Normalized Ratio; IQR, Interquartile Range; LASSO, Least Absolute Shrinkage and Selection Operator; LightGBM, Light Gradient Boosting Machine; LR, Logistic Regression; MAP, Mean Arterial Pressure; MAPE, Mean Absolute Prediction Error; MIMIC-IV, Medical Information Mart for Intensive Care IV; ML, Machine Learning; MV, Mechanical Ventilation; NPV, Negative Predictive Value; PCO₂, Partial Pressure of Carbon Dioxide; PO₂, Partial Pressure of Oxygen; PPV, Positive Predictive Value; PT, Prothrombin Time; qSOFA, Quick Sequential Organ Failure Assessment; RBC, Red Blood Cell Count; RDW, Red Cell Distribution Width; RF, Random Forest; ROC, Receiver Operating Characteristic; RR, Respiratory Rate; SBP, Systolic Blood Pressure; SD, Standard Deviation; SHAP, SHapley Additive exPlanations; SIRS, Systemic Inflammatory Response Syndrome; SOFA, Sequential Organ Failure Assessment; SPO₂, Peripheral Oxygen Saturation; SVM, Support Vector Machine; T2DM, Type 2 Diabetes Mellitus; WBC, White Blood Cell Count; XGBoost, Extreme Gradient Boosting.

Declarations

Ethics approval and consent to participate

This secondary analysis used publicly available, de-identified MIMIC-IV and eICU data. No additional ethics approval or individual informed consent was required.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed in this study are available through PhysioNet: MIMIC-IV (version 3.1) and the eICU Collaborative Research Database. Access requires completion of the relevant credentialing process and acceptance of the data-use agreement.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partly supported by the Guangxi Science and Technology Base and Talent Special Project (Grant 2021AC19394), the National Natural Science Foundation of China (Grant 82000560), and the Committee of Science and Technology of Changning District, Shanghai (Grant CNKW2022Y59).

Authors' contributions

HW and YW conceptualized the study, performed the formal analysis, and drafted the manuscript. HL and XX curated and preprocessed the data. XS performed feature selection and model validation. ZC and ZJ supervised the study, acquired funding, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

Not applicable.

References

1. Birkeland KI, Bodegard J, Eriksson JW, et al. Heart failure and chronic kidney disease manifestation and mortality risk associations in type 2 diabetes: a large multinational cohort study. *Diabetes Obes Metab*. 2020;22:1607-18. doi:10.1111/dom.14074.
2. Critchley JA, Carey IM, Harris T, DeWilde S, Hosking FJ, Cook DG. Glycemic control and risk of infections among people with type 1 or type 2 diabetes in a large primary care cohort study. *Diabetes Care*. 2018;41:2127-35. doi:10.2337/dc18-0287.
3. Liu YC, Yao Y, Yu MM, Gao YL, Qi AL, Jiang TY, et al. Frequency and mortality of sepsis and septic shock in China: a systematic review and meta-analysis. *BMC Infect Dis*. 2022;22:564. doi:10.1186/s12879-022-07543-8.
4. Paoli CJ, Reynolds MA, Sinha M, et al. Epidemiology and costs of sepsis in the United States: an analysis based on timing of diagnosis and severity level. *Crit Care Med*. 2018;46:1889-97. doi:10.1097/CCM.0000000000003342.
5. Wang Y, et al. Unveiling shared diagnostic biomarkers and molecular mechanisms between T2DM and sepsis: insights from bioinformatics to experimental assays. *FASEB J*. 2024;38:e70104. doi:10.1096/fj.202401872R.
6. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315:762-74. doi:10.1001/jama.2016.0288.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347-58. doi:10.1056/NEJMra1814259.
8. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. 2020;46:383-400. doi:10.1007/s00134-019-05872-y.
9. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One*. 2017;12:e0174708. doi:10.1371/journal.pone.0174708.
10. Negassa A, Ahmed S, Zolty R, Patel SR. Prediction model using machine learning for mortality in patients with heart failure. *Am J Cardiol*. 2021;153:86-93. doi:10.1016/j.amjcard.2021.05.044.
11. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail*. 2021;8:106-15. doi:10.1002/ehf2.13073.
12. Wändell P, Carlsson AC, Eriksson J, Wachtler C, Ruge T. A machine learning tool for identifying newly diagnosed heart failure in individuals with known diabetes in primary care. *ESC Heart Fail*. 2025;12:613-21. doi:10.1002/ehf2.15115.
13. Kwiendacz H, Huang B, Chen Y, et al. Predicting major adverse cardiac events in diabetes and chronic kidney disease: a machine learning study from the Silesia Diabetes-Heart Project. *Cardiovasc Diabetol*. 2025;24:21. doi:10.1186/s12933-025-02615-w.
14. Wang R, Dunzhu Z, Tan Z, et al. Development and validation of an interpretable machine learning model for early prediction in patients with diabetes and sepsis. *Sci Rep*. 2025;15:38546. doi:10.1038/s41598-025-22488-z.
15. McDermott MBA, Wang S, Marinsek N, et al. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med*. 2021;13:eabb1655. doi:10.1126/scitranslmed.abb1655.
16. Wang H, et al. Predictive modeling using artificial intelligence and machine learning algorithms on electronic health record data: advantages and challenges. *Crit Care Clin*. 2023;39:647-73. doi:10.1016/j.ccc.2023.02.001.
17. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315:801-10. doi:10.1001/jama.2016.0287.
18. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179:764-74. doi:10.1093/aje/kwt312.
19. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58:267-88. doi:10.1111/j.2517-6161.1996.tb02080.x.
20. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36:1-13. doi:10.18637/jss.v036.i11.
21. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565-74. doi:10.1177/0272989X06295361.
22. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *arXiv*. 2017;arXiv:1705.07874. doi:10.48550/arXiv.1705.07874.
23. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441.
24. Moser J, van der Aart TJ, Bouma HR. Urinary proteomics in sepsis-associated AKI. *Crit Care*. 2025;29:77. doi:10.1186/s13054-025-05306-w.
25. Kuwabara S, Goggins E, Okusa MD. The pathophysiology of sepsis-associated AKI. *Clin J Am Soc Nephrol*. 2022;17:1050-69. doi:10.2215/CJN.00850122.

26. Humes HD, Goldstein SL, Yessayan LT, et al. Safety summary of the selective cytopheretic device: a review of safety data across multiple clinical trials in ICU patients with acute kidney injury and multiple organ failure. *Crit Care Explor.* 2023;5:e0995. doi:10.1097/CCE.0000000000000995.
27. Zakyntinos GE, Tsolaki V, Xanthopoulos A, et al. Prevalence, risk factors, and mortality of new-onset atrial fibrillation in mechanically ventilated critically ill patients. *J Clin Med.* 2024;13:6750. doi:10.3390/jcm13226750.
28. Nseir S, Zerimech F, Jaillette E, et al. Microaspiration in intubated critically ill patients: diagnosis and prevention. *Infect Disord Drug Targets.* 2011;11:413-23. doi:10.2174/187152611796504827.
29. Luo Y, Wu Z, Wu J, et al. Revisiting ventilator-induced lung injury: from mechanical power to immunometabolic interaction. *Front Immunol.* 2026;17:1754983. doi:10.3389/fimmu.2026.1754983.
30. Cerny J, Rosmarin AG. Why does my patient have leukocytosis? *Hematol Oncol Clin North Am.* 2012;26:303-19, viii. doi:10.1016/j.hoc.2012.01.001.
31. Di Somma S, Travaglino F, Salerno G, et al. Comparison between white blood cell count, procalcitonin, and C-reactive protein as diagnostic and prognostic biomarkers of infection or sepsis in patients presenting to the emergency department. *Clin Chem Lab Med.* 2014;52:1465-72. doi:10.1515/ccdm-2014-0210.
32. Nierhaus A, Klatte S, Linssen J, et al. Revisiting the white blood cell count: immature granulocytes count as a diagnostic marker to discriminate between SIRS and sepsis: a prospective, observational study. *BMC Immunol.* 2013;14:8. doi:10.1186/1471-2172-14-8.
33. Col-Araz N, Pehlivan S, Baspinar O, et al. Role of cytokine gene (IFN- γ , TNF- α , TGF- β 1, IL-6, and IL-10) polymorphisms in the pathogenesis of acute rheumatic fever in Turkish children. *Eur J Pediatr.* 2012;171:1103-8. doi:10.1007/s00431-012-1689-5.
34. Zhao Y, Zhang B. Association between body temperature and all-cause mortality in patients with sepsis: analysis of the MIMIC-IV database. *Eur J Med Res.* 2024;29:630. doi:10.1186/s40001-024-02219-2.
35. Niven DJ, Laupland KB, et al. Diagnosis and management of temperature abnormality in ICUs: a EURO-BACT investigators survey. *Crit Care.* 2013;17:R289. doi:10.1186/cc13153.
36. Prachanukool T, Sanguanwit P, Suttapanit K, Yuksen C, et al. Initial venous lactate levels as a predictor of mortality in severe sepsis: a single-center retrospective cohort study. *World J Emerg Med.* 2022;13:396-9. doi:10.5847/wjem.j.1920-8642.2022.078.
37. Müller J, Radej J, Horak J, et al. Lactate: the fallacy of oversimplification. *Biomedicines.* 2023;11:3192. doi:10.3390/biomedicines11123192.
38. Bakker J. Lactate is the target for early resuscitation in sepsis. *Rev Bras Ter Intensiva.* 2017;29:124-7. doi:10.5935/0103-507X.20170021.
39. Ito H, Hosomi S, Koyama Y, et al. Sepsis-associated encephalopathy: a mini-review of inflammation in the brain and body. *Front Aging Neurosci.* 2022;14:912866. doi:10.3389/fnagi.2022.912866.
40. Borges-Sa M, Giglio A, Aranda M, et al. Decoding sepsis: a 16-year retrospective analysis of activation patterns, mortality predictors, and outcomes from a hospital-wide sepsis protocol. *J Clin Med.* 2025;14:5759. doi:10.3390/jcm14165759.
41. Wei W, Wei W, Fang D, et al. The impact of sedation and analgesia scores on prognosis in critically ill sepsis patients with sepsis-associated encephalopathy: a retrospective analysis. *Front Neurol.* 2025;16:1622964. doi:10.3389/fneur.2025.1622964.
42. Ganesh K, Sharma A. A profile of metabolic acidosis in patients with sepsis in an intensive care unit setting. *Int J Crit Illn Inj Sci.* 2016;6:178-81. doi:10.4103/2229-5151.195417.
43. Weemaes M, Hiele M, Vermeersch P. High anion gap metabolic acidosis caused by D-lactate: mind the time of blood collection. *Biochem Med (Zagreb).* 2020;30:011001. doi:10.11613/BM.2020.011001.
44. Hou Y, Li R, Li J, et al. Association between changes in corrected anion gap and mortality among critically ill patients during ICU stay: a multicenter observational study. *Front Physiol.* 2025;16:1469985. doi:10.3389/fphys.2025.1469985.

Figures

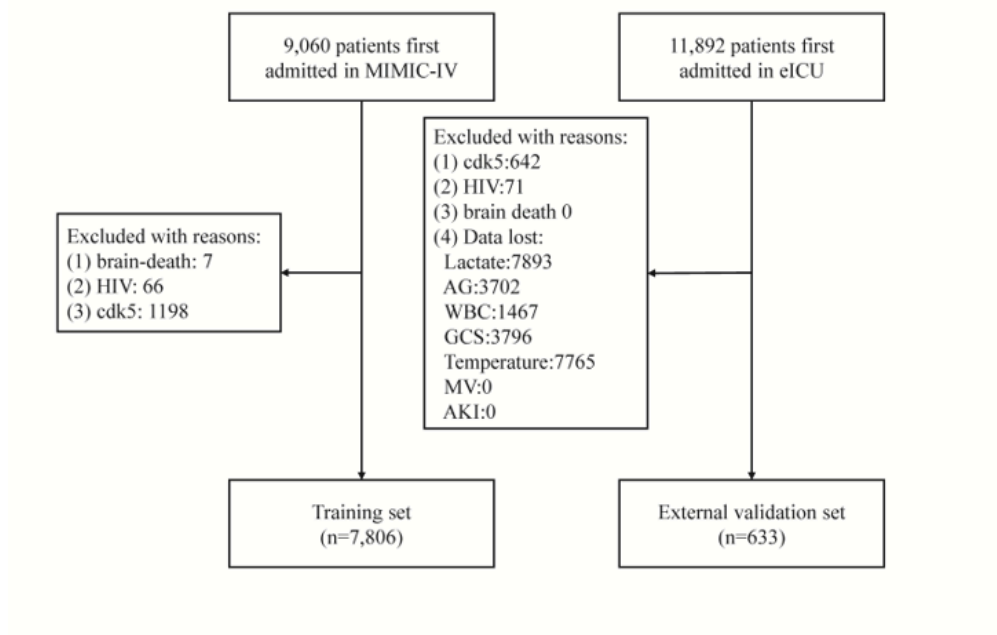


Figure 1
Dataset construction.

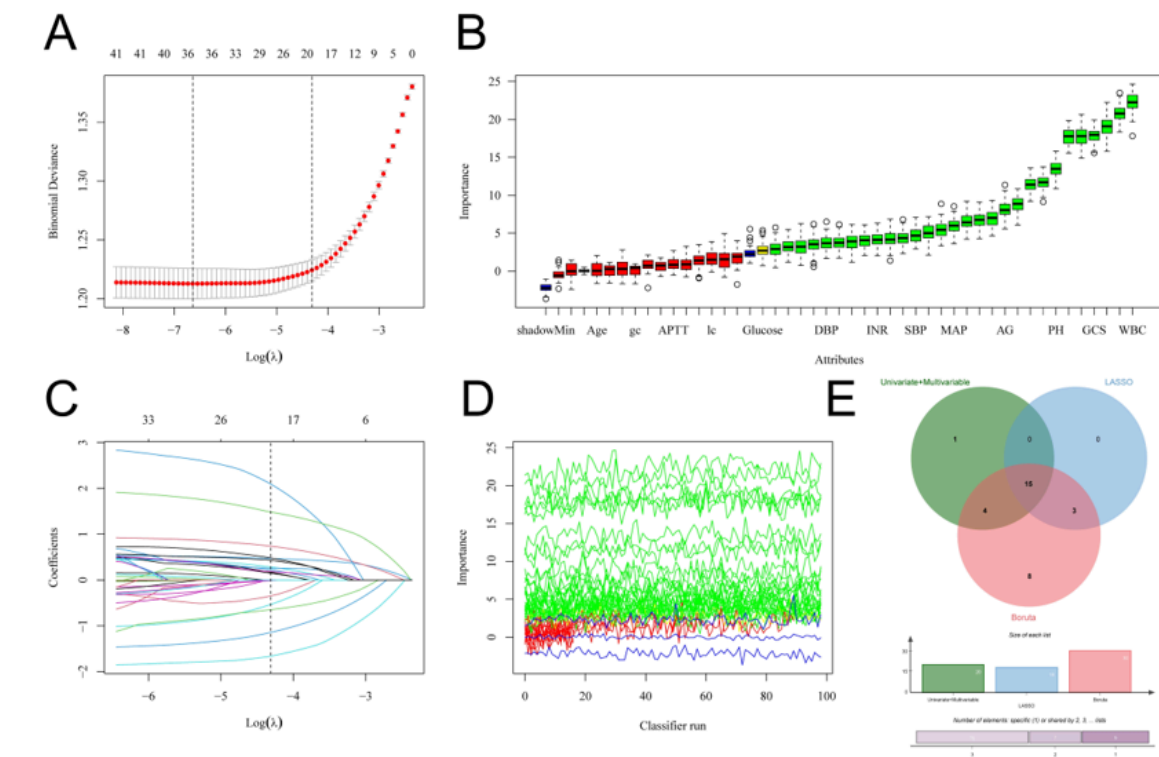


Figure 2
Feature-selection plots. (A, C) Binomial deviance and coefficient paths from least absolute shrinkage and selection operator regression; (B, D) Boruta feature importance; (E) overlap among the three feature-selection methods.

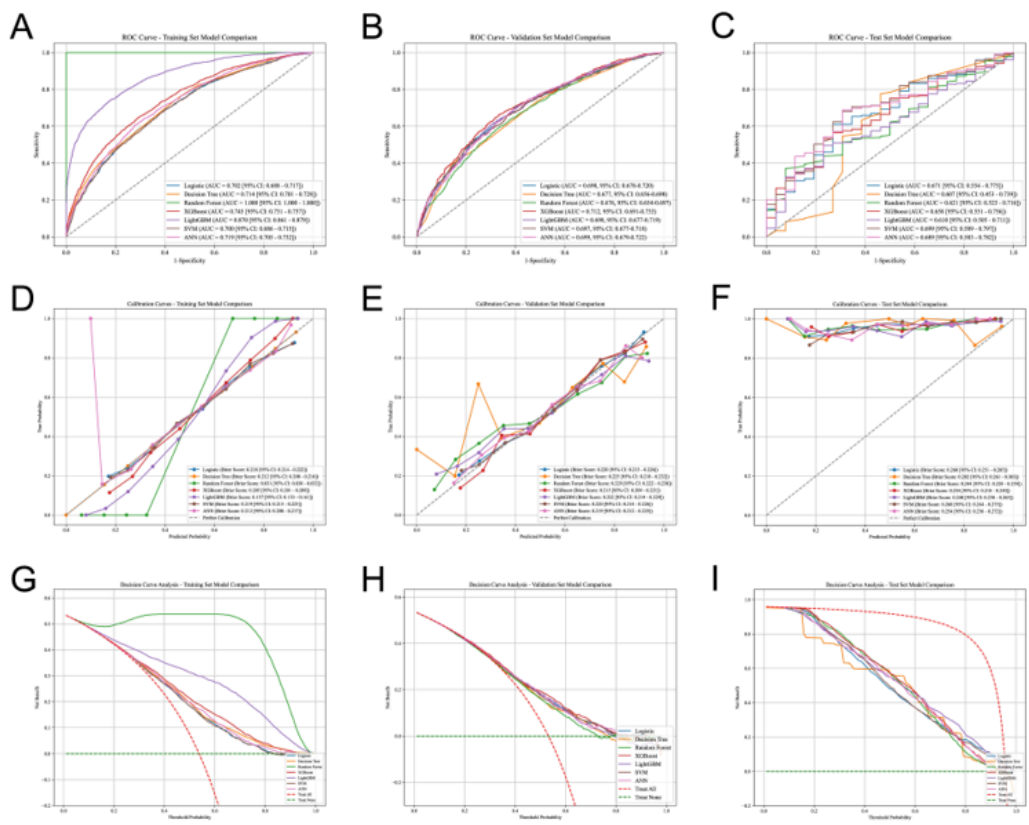


Figure 3

Model performance. Receiver operating characteristic, calibration, and decision curves are shown for the training set (A, D, G), test set (B, E, H), and external validation set (C, F, I).

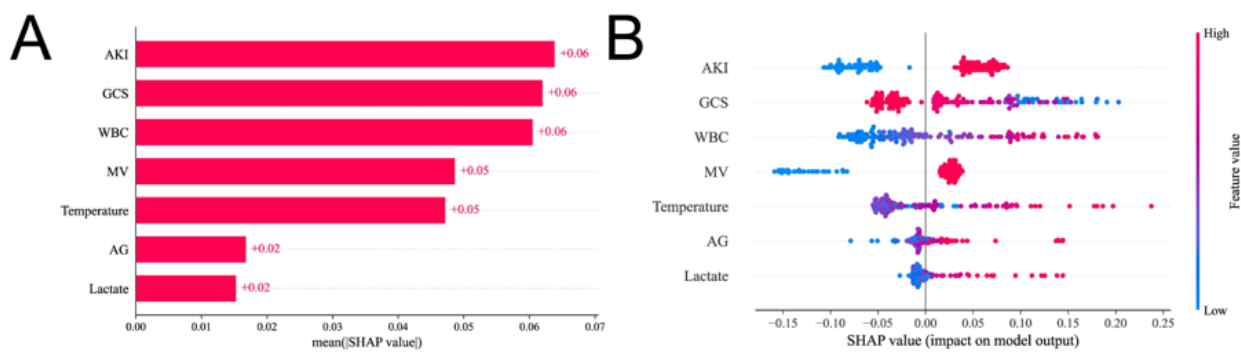


Figure 4

Global SHAP interpretation. (A) Feature importance ranked by mean absolute SHAP value; (B) beeswarm plot of SHAP values for the seven predictors.

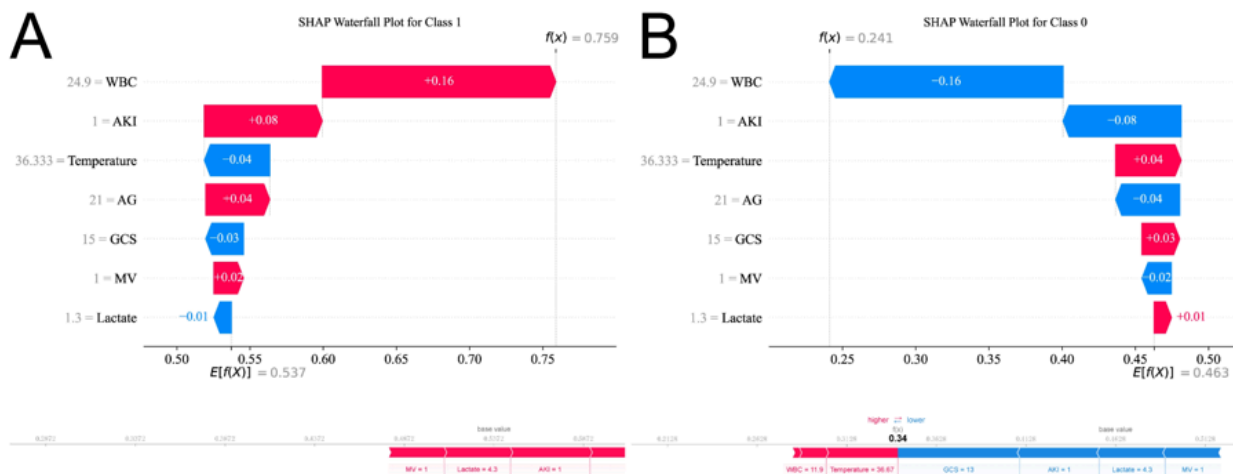


Figure 5
Local SHAP interpretations for representative patients. (A) Contributions for a patient with higher predicted sepsis risk; (B) contributions for a patient with lower predicted risk.

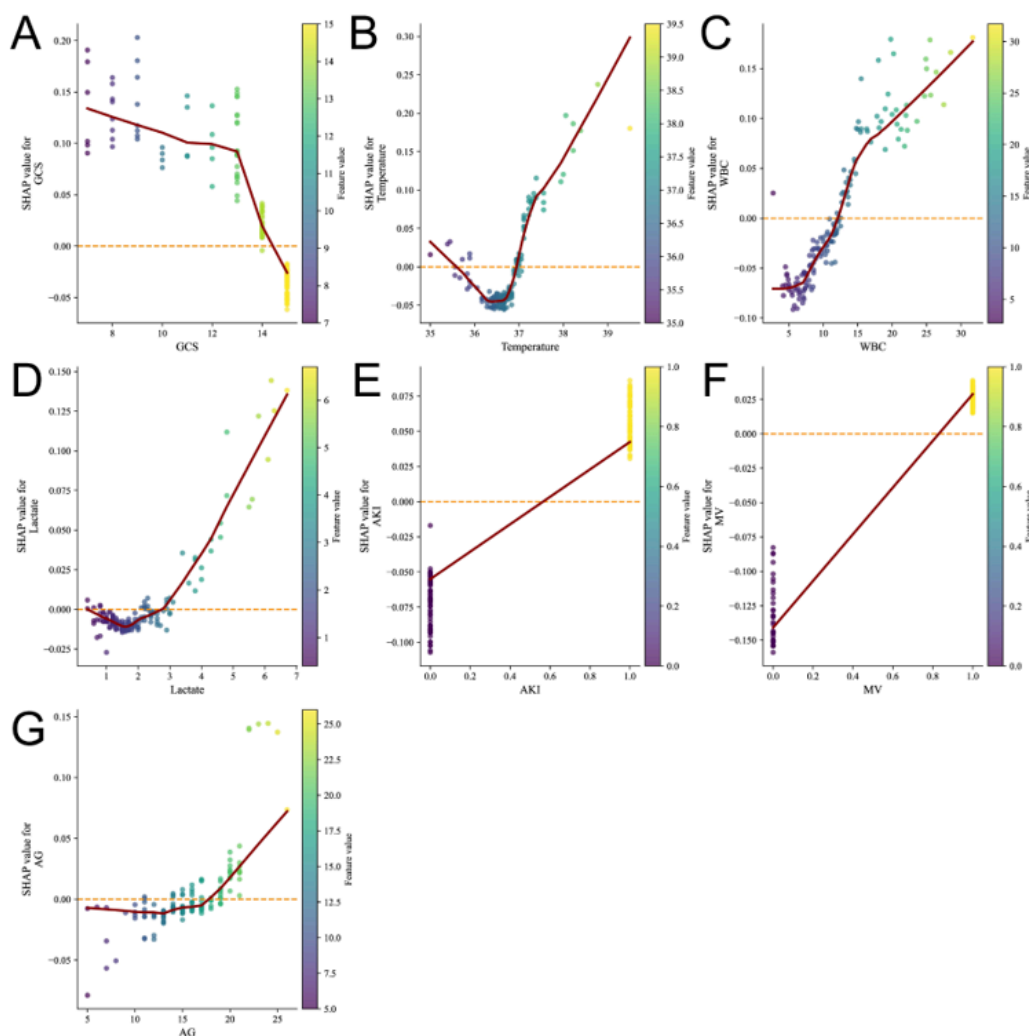


Figure 6
SHAP dependence plots showing the association between each core predictor and its SHAP value.