Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Machine Learning Model for Predicting In-Hospital Mortality among ICU Patients with Cirrhosis and Sepsis

Hui Chen
chenhui_ella@163.com

Huadong Hospital
Chen Yang
Shenzhen Second People's Hospital
Jiahao Zhang
Dyson Institute of Engineering and Technology
Yiping Dai
Zhongnan Hospital of Wuhan University
Pangaozhi Mo
Zhongnan Hospital of Wuhan University
Zhiyong Peng
Zhongnan Hospital of Wuhan University
Zhixiong Wu
Huadong Hospital

Research Article

Keywords: machine learning, sepsis, liver cirrhosis, mortality prediction model

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10717154/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

To develop and internally and externally validate a model for predicting in-hospital mortality risk among patients with cirrhosis and sepsis, using routine clinical data from the first 24 hours after ICU admission.

Methods

This was a retrospective cohort study. The development cohort was drawn from MIMIC-IV (2008–2019; n = 2463) and randomly split 8:2 into a training set (n = 1970) and an internal validation set (n = 493). The external validation cohort comprised 237 patients from Zhongnan Hospital of Wuhan University (2015–2025). Adults with a diagnosis of cirrhosis who met Sepsis-3 criteria were included; patients with an ICU length of stay ≤ 24 hours or critical missing data were excluded. The primary endpoint was all-cause in-hospital mortality. Seven algorithms were trained: Logistic Regression, Ridge, Lasso, Decision Tree, Random Forest, XGBoost, and LightGBM. Missing values were handled with IterativeImputer. The primary criterion for selecting the optimal model was the arithmetic mean of AUROCs across cross-validation, internal validation, and external validation. Accuracy, precision, recall/sensitivity, specificity, F1 score, and Brier score were also reported. Performance was compared with conventional scores including SOFA, SAPS II, OASIS, LODS, MELD, and ABIC. The final model was interpreted with SHAP.

Results

A total of 2700 patients were included (2463 from MIMIC-IV and 237 from Zhongnan Hospital). XGBoost showed the best overall performance, with AUROCs of 0.940 (training), 0.833 (internal validation), and 0.776 (external validation); the three-set mean AUROC was 0.814. Discrimination of XGBoost exceeded that of the compared conventional scores (DeLong test, $P < 0.05$). The leading SHAP contributors included liver transplantation status, total bilirubin, prothrombin time, bloodstream infection, and urine output.

Conclusions

An XGBoost model based on first-day ICU data can stratify in-hospital mortality risk in patients with cirrhosis and sepsis, with discrimination superior to conventional scores and with clinically interpretable predictors.

Background

Cirrhosis represents the end stage of various chronic liver diseases, including hepatitis B and C virus infection, steatotic liver disease, alcohol-related liver disease, and autoimmune hepatitis, and its complications seriously threaten survival[1, 2]. Liver disease currently ranks as the 11th leading cause of death worldwide; an estimated two million liver-related deaths occur each year, accounting for approximately 4% of all deaths, with an age-standardized death rate of about 16.6 per 100,000 population[3, 4]. Hospitalized patients with cirrhosis face high morbidity and mortality: ICU admission rates range from 20% to 40%, ICU mortality from 34% to 72%, and mortality is about 1.5-fold higher than in patients without cirrhosis[5, 6]. Bacterial infection is a common clinical complication in cirrhosis; contemporary global data indicate that approximately 30% to 35% of hospitalized patients with cirrhosis have bacterial infection at or shortly after admission, and prognosis is often poor[7]. Infections markedly worsen short-term prognosis in cirrhosis—classic multicenter data indicate that infection increases mortality approximately fourfold[9]—and studies of acute decompensation have further underscored the prognostic impact of infection and organ failure[8]. Early identification of high-risk patients and timely intervention are therefore essential to improve outcomes.

Patients with cirrhosis have a markedly increased risk of infection and progression to sepsis, with short-term mortality more than fourfold higher than that of patients without liver disease[10, 11]. This elevated risk arises mainly from the following pathophysiologic mechanisms[12]. First, cirrhosis-associated immune dysfunction (CAID) is common and characterized by leukopenia, impaired synthesis of innate immune proteins, immune-cell exhaustion, reduced neutrophil phagocytosis and oxidative burst, and overactivation of the tryptophan–kynurenine pathway. The coexistence of persistent systemic inflammation and defective host defense increases susceptibility to pathogens and impairs clearance, thereby promoting dissemination of infection and

multiorgan dysfunction[13]. Second, portal hypertension disrupts the intestinal barrier and increases permeability, facilitating translocation of gut bacteria into the systemic circulation; gut dysbiosis further amplifies this process in a vicious cycle[14, 15]. Third, the burden of antimicrobial resistance is substantial: repeated hospitalizations are associated with a high proportion of multidrug-resistant infections and an increased risk of failure of initial empiric therapy[16]. Together, these factors predispose patients with cirrhosis to severe infection, which may precipitate acute kidney injury, hepatic encephalopathy, and other serious complications and substantially increase mortality risk[17].

According to the Third International Consensus Definitions for Sepsis and Septic Shock, sepsis is life-threatening organ dysfunction caused by infection[18]. Cirrhosis has been identified as an independent risk factor for death in septic shock[20], and concurrent cirrhosis in patients with sepsis substantially increases short- and long-term mortality. Guidelines emphasize that the early period after sepsis recognition is a critical intervention window; rapid identification of high-risk patients after ICU admission is closely linked to timely treatment and improved outcomes[19, 21]. However, the SOFA score was not designed for early risk assessment specifically in ICU patients with cirrhosis and sepsis. Although widely used in sepsis, baseline hepatic abnormalities such as elevated bilirubin and coagulopathy may be chronic in cirrhosis, limiting the sensitivity of SOFA for detecting acute deterioration and constraining its diagnostic utility[22, 23]. Delayed recognition of sepsis and delayed empiric antibiotics in this population may further worsen prognosis.

In recent years, machine learning has shown substantial potential for medical risk prediction. Unlike conventional scores that rely on preset weights and linear relationships, machine-learning algorithms can handle high-dimensional, nonlinear clinical data and automatically capture complex interactions among variables, which may enable more accurate prediction[24, 25]. Their capacity for feature selection may also help identify the most informative predictors among many candidate variables and provide new insights for clinical monitoring and mechanistic research.

Accordingly, this study aimed to use machine-learning methods in accordance with TRIPOD + AI guidance[26] to develop an in-hospital mortality prediction model based on clinical data from the first 24 hours after ICU admission in patients with cirrhosis and sepsis, with the goal of enabling early and accurate estimation of in-hospital death risk, and to interpret the model with SHAP[27], thereby providing an improved tool for early risk stratification and intervention decisions.

Methods

Study design and data sources

This multicenter retrospective cohort study used MIMIC-IV version 2.0 for model development and internal validation and an independent cohort from Zhongnan Hospital of Wuhan University for external validation. MIMIC-IV includes ICU admissions from 2008 to 2019 at Beth Israel Deaconess Medical Center[28]. Access was obtained after required training (CITI certification number 57917614). The external cohort included adults with cirrhosis and sepsis who were first admitted to the ICU between 1 January 2015 and 31 December 2025. The study was approved by the institutional ethics committee of Zhongnan Hospital of Wuhan University (Approval No. 2026049k). All data were de-identified, and informed consent was waived.

Participants

Inclusion criteria were age ≥ 18 years; a confirmed diagnosis of cirrhosis (ICD-9/ICD-10 in MIMIC-IV; ICD-10 in the external cohort); and fulfillment of Sepsis-3 criteria at ICU admission (suspected infection with an increase in SOFA score of ≥ 2 from baseline). For patients with repeated ICU admissions, only the first admission was included.

Exclusion criteria were ICU length of stay ≤ 24 hours and critical missing data.

Outcome

The primary endpoint was all-cause in-hospital mortality.

Variables

Variables were restricted to measurements within the first 24 hours after ICU admission; when multiple values were available, the worst value was used. Collected data included demographics; comorbidities and cirrhosis-related complications; laboratory tests; clinical severity scores (GCS, SOFA, SAPS II, OASIS, LODS, MELD, and ABIC); vital signs and urine output; and early interventions (invasive mechanical ventilation, continuous renal replacement therapy, and vasoactive drugs). Infection sites were classified from ICD codes as pulmonary, urinary, bloodstream, or other.

Statistical analysis and modeling

Continuous variables were summarized according to distribution: normally distributed data as mean ± SD and non-normally distributed data as median [interquartile range], and compared with t tests or Mann–Whitney U tests. Categorical variables were summarized as n (%) and compared with chi-square or Fisher exact tests. Analyses were performed in Python 3.10 using scikit-learn, XGBoost, LightGBM, SHAP[27], SciPy, pandas, and numpy. Two-sided $P < 0.05$ was considered statistically significant.

Missing values were imputed separately within each dataset using IterativeImputer (a multivariate chained-equation approach implemented in scikit-learn). Variables with > 30% missingness were excluded. Continuous features were Z-score standardized using means and standard deviations from the training set.

Seven models were developed: logistic regression, ridge regression, lasso regression, decision tree, random forest, XGBoost, and LightGBM. Hyperparameters were optimized on the training set by 5-fold cross-validation. Out-of-fold predictions were used to estimate cross-validation performance. Models were then refit on the full training set and evaluated on the internal and external validation sets. The primary selection criterion was the mean AUROC across cross-validation, internal validation, and external validation. Secondary metrics included accuracy, precision, sensitivity, specificity, F1 score, and Brier score, and a calibration curve was plotted for the final model. Conventional scores (SOFA, SAPS II, OASIS, LODS, MELD, and ABIC) were entered as single predictors in unpenalized logistic regression, and AUROCs were compared with the final model using DeLong tests. For interpretability, SHAP values were computed for the final model[27] (TreeExplainer for tree-based models). Global importance, summary plots, and an individual explanation example were generated from up to 1000 randomly sampled training observations. Key hyperparameters are listed in Table 1.

Table 1
Key hyperparameters of each model

Model	Key hyperparameters
Logistic regression	C = 1.0, penalty=l2, solver=lbfgs, fit_intercept=True, max_iter = 2000, tol = 0.0001, random_state = 42
Ridge regression	C = 1.0, penalty=l2, solver=lbfgs, fit_intercept=True, max_iter = 2000, tol = 0.0001, random_state = 42
Lasso regression	C = 0.5, penalty=l1, solver=liblinear, fit_intercept=True, max_iter = 2000, tol = 0.0001, random_state = 42
Decision tree	max_depth = 6, criterion=gini, min_samples_leaf = 20, min_samples_split = 2, random_state = 42, splitter=best
Random forest	n_estimators = 300, max_depth = 8, min_samples_leaf = 20, criterion=gini, max_features=sqrt, bootstrap=True, random_state = 42
XGBoost	n_estimators = 300, max_depth = 3, learning_rate = 0.03, subsample = 0.8, colsample_bytree = 0.7, reg_alpha = 1.0, reg_lambda = 0.1, random_state = 42
LightGBM	n_estimators = 800, max_depth = 5, learning_rate = 0.01, subsample = 0.8, colsample_bytree = 0.7, reg_alpha = 1.0, reg_lambda = 10.0, random_state = 42

Results

Study population and baseline characteristics

Study flow and grouping

A total of 2700 patients with cirrhosis and sepsis were included from MIMIC-IV and the Zhongnan Hospital ICU (MIMIC cohort, n = 2463; Zhongnan cohort, n = 237). As shown in Fig. 1, the MIMIC cohort was randomly split 8:2 into a training set (n = 1970; 558 deaths, 28.3%) and an internal validation set (n = 493; 138 deaths, 28.0%); the Zhongnan cohort served as the external validation set (n = 237; 22 deaths, 9.3%). Sepsis on ICU day 1 was defined as suspected infection with an increase in SOFA of ≥ 2 from baseline.

Baseline characteristics of survivors and nonsurvivors

As shown in Table 2, patients who died in hospital already exhibited more severe pathophysiologic derangement at admission. Compared with survivors, nonsurvivors had higher severity scores (SOFA, SAPS II, LODS, OASIS, MELD, and ABIC; all $P < 0.01$), greater hemodynamic instability (faster heart rate and lower systolic, diastolic, and mean arterial pressures; $P < 0.01$), and more marked respiratory dysfunction (higher respiratory rate and lower oxygen saturation; $P < 0.001$). Laboratory findings indicated more severe hepatic and renal injury (higher total bilirubin and creatinine; $P < 0.001$), worse coagulopathy, and a stronger systemic inflammatory response (higher white-cell count; $P < 0.001$). Nonsurvivors also more frequently received invasive mechanical ventilation, vasoactive drugs, and renal replacement therapy (all $P < 0.001$). Patterns in the internal validation set were similar to those in the training set. In the external validation set, nonsurvivors remained significantly worse for some markers (including INR, PT, SOFA, hepatic failure, septic shock, and CRRT use; $P < 0.05$), whereas differences in age, sex, most scoring systems, WBC, creatinine, BUN, total bilirubin, blood pressure, and heart rate were not significant, likely related to the smaller sample size.

Table 2
Baseline characteristics of survivors and nonsurvivors across datasets

Variable	Train survivor (n = 1412)	Train nonsurvivor (n = 558)	P	Internal survivor (n = 355)	Internal nonsurvivor (n = 138)	P	External survivor (n = 215)	External nonsurvivor (n = 22)	P
Age (years)	58.00 (50.00–65.00)	59.00 (51.00–67.00)	0.06	58.00 (51.00–65.00)	58.00 (51.00–66.75)	0.89	58.00 (49.00–67.50)	59.50 (49.00–68.75)	0.59
Female sex	493 (34.9%)	207 (37.1%)	0.37	118 (33.2%)	50 (36.2%)	0.53	73 (34.0%)	7 (31.8%)	1.00
Weight (kg)	81.80 (69.67–96.83)	84.30 (70.70–101.50)	0.05	80.00 (68.00–96.60)	89.90 (73.55–102.60)	< 0.01	62.00 (56.81–70.00)	62.10 (56.69–70.00)	0.98
BMI (kg/m²)	27.99 (24.45–32.68)	29.48 (24.81–34.65)	0.08	27.51 (23.45–32.17)	31.28 (25.49–36.61)	0.55	21.79 (20.14–24.45)	22.01 (19.89–24.14)	0.98
SOFA score	4.00 (3.00–6.00)	6.00 (4.00–8.00)	< 0.01	4.00 (3.00–6.00)	6.00 (4.00–8.00)	< 0.01	8.00 (5.00–10.00)	0.69	
GCS score	15.00 (15.00–15.00)	15.00 (14.00–15.00)	< 0.01	15.00 (15.00–15.00)	15.00 (14.00–15.00)	< 0.01	15.00 (9.00–15.00)	11.87 (5.50–15.00)	0.07
SAPS II score	38.00 (30.00–48.00)	50.00 (41.00–62.00)	< 0.01	39.00 (30.00–48.00)	49.00 (39.25–62.00)	< 0.01	36.00 (27.00–46.50)	41.00 (27.50–54.00)	0.36
LODS score	6.00 (4.00–8.00)	8.00 (6.00–10.00)	< 0.01	6.00 (4.00–8.00)	8.00 (6.00–11.00)	< 0.01	6.00 (3.00–8.00)	7.50 (4.25–10.00)	0.13
OASIS score	33.00 (27.00–38.00)	38.00 (32.00–45.00)	< 0.01	33.00 (27.00–39.00)	38.00 (33.00–44.00)	< 0.01	28.00 (22.00–34.00)	26.50 (22.00–30.75)	0.36
MELD score	22.34 (15.71–29.15)	32.00 (25.31–38.17)	< 0.01	23.00 (16.72–29.17)	31.35 (23.89–39.15)	< 0.01	18.40 (14.00–25.50)	19.97 (17.10–24.75)	0.13
ABIC score	8.31 (7.43–9.42)	9.65 (8.58–10.94)	< 0.01	8.47 (7.54–9.40)	9.64 (8.39–10.64)	< 0.01	8.13 (7.03–9.33)	8.88 (7.22–9.63)	0.29
Hypertension	537 (38.0%)	181 (32.4%)	0.02	130 (36.6%)	48 (34.8%)	0.75	31 (14.4%)	7 (31.8%)	0.06
Atherosclerotic heart disease	176 (12.5%)	60 (10.8%)	0.32	45 (12.7%)	19 (13.8%)	0.77	19 (8.8%)	1 (4.5%)	0.70
Atrial fibrillation	248 (17.6%)	153 (27.4%)	< 0.01	70 (19.7%)	35 (25.4%)	0.18	13 (6.0%)	2 (9.1%)	0.64
Diabetes mellitus	467 (33.1%)	147 (26.3%)	< 0.0136	121 (34.1%)	45 (32.6%)	0.83	42 (19.5%)	5 (22.7%)	0.78
Chronic kidney disease	298 (21.1%)	132 (23.7%)	0.23	75 (21.1%)	30 (21.7%)	0.90	23 (10.7%)	6 (27.3%)	0.04
Chronic viral hepatitis	268 (19.0%)	68 (12.2%)	< 0.01	66 (18.6%)	21 (15.2%)	0.43	208 (96.7%)	22 (100.0%)	1.00
Alcoholic hepatitis	739 (52.3%)	334 (59.9%)	< 0.01	194 (54.6%)	76 (55.1%)	1.00	26 (12.1%)	1 (4.5%)	0.48

Variable	Train survivor (n = 1412)	Train nonsurvivor (n = 558)	P	Internal survivor (n = 355)	Internal nonsurvivor (n = 138)	P	External survivor (n = 215)	External nonsurvivor (n = 22)	P
Drug-induced liver injury	6 (0.4%)	6 (1.1%)	0.11	1 (0.3%)	0 (0.0%)	1.00	6 (2.8%)	1 (4.5%)	0.50
Cholestatic liver disease	43 (3.0%)	11 (2.0%)	0.22	8 (2.3%)	2 (1.4%)	0.73	25 (11.6%)	0 (0.0%)	0.14
Autoimmune hepatitis	14 (1.0%)	7 (1.3%)	0.63	4 (1.1%)	0 (0.0%)	0.58	15 (7.0%)	0 (0.0%)	0.37
Hepatic encephalopathy	138 (9.8%)	63 (11.3%)	0.32	33 (9.3%)	16 (11.6%)	0.50	30 (14.0%)	6 (27.3%)	0.12
Esophagogastric varices	406 (28.8%)	113 (20.3%)	< 0.01	106 (29.9%)	31 (22.5%)	0.12	118 (54.9%)	14 (63.6%)	0.50
Gastrointestinal bleeding	104 (7.4%)	23 (4.1%)	0.01	27 (7.6%)	8 (5.8%)	0.56	54 (25.1%)	6 (27.3%)	0.80
Hepatic failure	251 (17.8%)	170 (30.5%)	< 0.01	73 (20.6%)	52 (37.7%)	< 0.01	35 (16.3%)	7 (31.8%)	0.08
Hepatorenal syndrome	159 (11.3%)	143 (25.6%)	< 0.01	39 (11.0%)	34 (24.6%)	< 0.01	13 (6.0%)	1 (4.5%)	1.00
Acute liver failure	0 (0.0%)	1.00	0 (0.0%)	1.00	8 (3.7%)	1 (4.5%)	0.59		
Peritonitis	95 (6.7%)	52 (9.3%)	0.06	26 (7.3%)	15 (10.9%)	0.21	19 (8.8%)	3 (13.6%)	0.44
Ascites	381 (27.0%)	179 (32.1%)	0.03	105 (29.6%)	42 (30.4%)	0.91	90 (41.9%)	8 (36.4%)	0.66
Septic shock	252 (17.8%)	172 (30.8%)	< 0.01	55 (15.5%)	44 (31.9%)	< 0.01	27 (12.6%)	8 (36.4%)	< 0.01
Heart rate (beats/min)	104.00 (91.00–117.00)	112.00 (99.00–124.00)	< 0.01	104.00 (89.00–118.00)	107.50 (95.25–123.75)	0.02	110.00 (96.00–126.50)	109.00 (90.75–125.50)	0.62
SBP (mmHg)	88.00 (80.00–97.00)	83.00 (75.00–91.00)	< 0.01	87.00 (79.00–97.00)	82.00 (73.00–92.00)	< 0.01	85.00 (66.00–98.75)	73.50 (63.25–96.50)	0.26
DBP (mmHg)	46.00 (40.00–52.00)	43.00 (36.00–48.00)	< 0.01	45.00 (39.00–51.75)	41.00 (35.00–46.00)	< 0.01	43.00 (34.00–52.00)	40.00 (32.00–49.75)	0.42
MAP (mmHg)	58.00 (51.00–65.00)	55.00 (49.00–61.00)	< 0.01	57.00 (50.50–63.00)	53.50 (48.00–61.00)	0.01	56.00 (46.00–68.00)	49.50 (44.68–65.75)	0.48
Respiratory rate (breaths/min)	27.00 (23.00–31.00)	29.00 (25.00–34.00)	< 0.01	26.00 (23.00–31.00)	28.75 (25.00–33.00)	< 0.01	29.00 (26.00–30.00)	27.50 (23.00–30.00)	0.13
Temperature (°C)	37.22 (37.00–37.67)	37.17 (36.89–37.56)	< 0.01	37.22 (36.94–37.58)	37.06 (36.78–37.56)	0.03	36.95 (36.70–37.20)	36.60 (36.37–36.80)	< 0.01
SpO_2 (%)	93.00 (91.00–95.00)	92.00 (89.00–94.00)	< 0.01	93.00 (91.00–95.00)	92.00 (88.00–94.00)	< 0.01	91.00 (84.00–95.00)	91.00 (77.50–96.50)	0.94
Urine output (mL)	1272.50 (769.75–1936.25)	752.00 (292.75–1343.21)	< 0.01	1260.00 (695.00–1977.50)	866.50 (258.00–1415.59)	< 0.01	965.35 (625.00–1229.75)	580.00 (112.12–1019.86)	0.01

Variable	Train survivor (n = 1412)	Train nonsurvivor (n = 558)	P	Internal survivor (n = 355)	Internal nonsurvivor (n = 138)	P	External survivor (n = 215)	External nonsurvivor (n = 22)	P
Glucose (mmol/L)	9.17 (7.11–13.17)	8.72 (7.11–11.49)	0.02	8.94 (6.89–12.47)	9.28 (7.57–11.43)	0.69	10.08 (7.35–12.50)	8.60 (5.30–10.31)	0.04
WBC (×10⁹/L)	11.70 (7.70–16.92)	15.00 (9.90–21.20)	< 0.01	12.30 (8.20–18.45)	15.35 (9.68–20.77)	< 0.01	11.00 (6.38–14.14)	11.39 (8.26–12.95)	0.88
Hemoglobin (g/L)	87 (73–101)	80 (69–96.7)	< 0.01	85 (73–99.5)	81 (68.3–95)	0.16	83.22 (65.00–99.00)	78.00 (64.25–91.56)	0.49
Hematocrit max (%)	31.00 (27.60–35.30)	29.40 (25.90–34.90)	< 0.01	30.50 (26.80–35.00)	30.45 (25.75–34.72)	0.56	27.65 (22.55–32.25)	26.33 (22.47–29.71)	0.62
Hematocrit min (%)	25.80 (22.30–30.00)	24.30 (20.90–29.00)	< 0.01	25.80 (21.95–29.45)	24.50 (21.00–29.27)	0.33	25.00 (19.80–28.75)	24.01 (19.77–25.70)	0.53
Platelets (×10⁹/L)	83.00 (52.75–133.00)	78.50 (47.00–133.00)	0.20	90.00 (56.50–134.50)	69.50 (45.25–111.00)	0.01	86.00 (55.00–118.50)	72.50 (55.00–99.02)	0.47
INR	1.70 (1.40–2.20)	2.30 (1.80–3.10)	< 0.01	1.70 (1.40–2.10)	2.30 (1.70–3.20)	< 0.01	1.62 (1.31–1.98)	2.04 (1.85–2.47)	< 0.01
PT (s)	18.60 (15.40–24.00)	24.60 (19.42–33.25)	< 0.01	18.60 (15.75–23.20)	24.60 (18.45–33.60)	< 0.01	17.70 (14.35–21.60)	22.09 (20.09–26.82)	< 0.01
PTT (s)	38.50 (32.48–51.73)	48.90 (37.30–68.33)	< 0.01	39.50 (32.70–52.30)	50.45 (38.38–68.78)	< 0.01	17.70 (14.35–21.60)	22.09 (20.09–26.82)	< 0.01
Total bilirubin (μmol/L)	42.75 (18.81–97.81)	105.2 (39.33–288.1)	< 0.01	47.88 (22.23–102.6)	90.63 (39.67–197.8)	< 0.01	42.20 (20.00–93.16)	61.65 (30.75–110.35)	0.19
Albumin (g/L)	28 (25–32)	28 (23–32)	< 0.01	29 (25.8–33)	27 (22.6–32.7)	0.01	28.50 (25.50–32.08)	27.71 (22.98–31.19)	0.39
ALT (U/L)	41.00 (23.00–125.34)	44.00 (25.00–111.00)	0.38	38.00 (20.00–84.65)	38.00 (24.00–80.75)	0.64	57.00 (21.00–176.78)	91.00 (25.00–141.82)	0.88
AST (U/L)	86.00 (45.00–249.75)	111.00 (55.25–287.50)	< 0.01	78.00 (39.00–176.00)	84.50 (52.25–186.50)	0.12	107.00 (42.00–334.73)	168.00 (55.00–292.01)	0.91
BUN (mmol/L)	10.36 (6.43–17.14)	15 (8.21–23.57)	< 0.01	10.36 (6.43–17.32)	13.93 (7.59–20.36)	0.01	9.58 (6.45–14.41)	9.39 (7.70–14.55)	0.72
Creatinine (μmol/L)	114.9 (79.56–185.6)	185.6 (114.9–316.5)	< 0.01	123.8 (79.56–203.3)	176.8 (106.1–291.7)	< 0.01	99.80 (68.55–190.25)	97.44 (56.18–138.62)	0.13
Sodium (mmol/L)	139.00 (136.00–142.00)	138.00 (133.00–142.00)	< 0.01	139.00 (135.00–142.00)	139.00 (135.00–143.75)	0.18	139.00 (136.00–143.27)	140.70 (137.93–145.80)	0.16
Potassium (mmol/L)	4.50 (4.00–5.10)	4.65 (4.10–5.40)	< 0.01	4.40 (4.00–5.10)	4.70 (4.12–5.30)	0.02	4.19 (3.67–4.61)	4.45 (4.13–5.27)	0.04

Variable	Train survivor (n = 1412)	Train nonsurvivor (n = 558)	P	Internal survivor (n = 355)	Internal nonsurvivor (n = 138)	P	External survivor (n = 215)	External nonsurvivor (n = 22)	P
Calcium (mmol/L)	2.12 (2–2.25)	2.17 (2.02–2.33)	< 0.01	2.12 (2–2.25)	2.15 (2–2.35)	0.16	2.03 (1.91–2.20)	2.02 (1.90–2.27)	0.91
Chloride min (mmol/L)	101.00 (97.00–106.00)	98.00 (92.00–103.00)	< 0.01	100.00 (96.00–105.00)	99.00 (95.00–105.00)	0.25	104.84 (101.05–109.10)	107.10 (101.80–109.57)	0.47
Chloride max (mmol/L)	105.00 (101.00–110.00)	102.00 (97.00–108.00)	< 0.01	105.00 (100.00–109.00)	104.00 (99.00–109.00)	0.41	106.60 (103.95–110.60)	108.10 (103.25–111.30)	0.49
MCH (pg)	31.64 (29.50–33.70)	32.35 (30.10–34.80)	< 0.01	31.30 (29.15–33.40)	32.30 (30.50–34.58)	< 0.01	31.47 (29.71–32.95)	32.15 (31.05–34.38)	0.05
MCHC (g/L)	332 (321–341)	331 (320–342)	0.72	330 (319–340)	327.5 (318–338.8)	0.31	31.47 (29.71–32.95)	32.15 (31.05–34.38)	0.05
MCV (fL)	95.00 (90.00–100.00)	97.00 (92.00–103.75)	< 0.01	94.00 (89.00–100.00)	98.00 (93.00–104.00)	< 0.01	93.50 (88.30–99.03)	97.06 (92.21–101.20)	0.05
RDW (%)	16.30 (14.90–18.20)	17.20 (15.43–19.58)	< 0.01	16.70 (14.95–18.50)	17.55 (16.00–19.68)	< 0.01	17.50 (15.75–18.70)	17.97 (16.62–19.60)	0.26
Lactate (mmol/L)	2.99 (1.90–4.38)	3.94 (2.50–6.26)	< 0.01	3.00 (1.85–4.50)	4.23 (2.54–6.60)	< 0.01	3.80 (1.90–5.54)	5.20 (2.73–6.85)	0.15
CRRT	158 (11.2%)	184 (33.0%)	< 0.01	38 (10.7%)	49 (35.5%)	< 0.01	6 (2.8%)	3 (13.6%)	0.04
Invasive mechanical ventilation	703 (49.8%)	317 (56.8%)	0.01	176 (49.6%)	83 (60.1%)	0.04	51 (23.7%)	0 (0.0%)	< 0.01
Liver transplantation	202 (14.3%)	3 (0.5%)	< 0.01	47 (13.2%)	0 (0.0%)	< 0.01	42 (19.5%)	0 (0.0%)	0.02
Pulmonary infection	285 (20.2%)	177 (31.7%)	< 0.01	74 (20.8%)	42 (30.4%)	0.03	112 (52.1%)	12 (54.5%)	1.00
Urinary tract infection	10 (0.7%)	4 (0.7%)	1.00	9 (2.5%)	0 (0.0%)	0.07	15 (7.0%)	0 (0.0%)	0.37
Bloodstream infection	405 (28.7%)	339 (60.8%)	< 0.01	113 (31.8%)	78 (56.5%)	< 0.01	27 (12.6%)	3 (13.6%)	0.75
28-day mortality	102 (7.2%)	497 (89.1%)	< 0.01	26 (7.3%)	124 (89.9%)	< 0.01			
90-day mortality	241 (17.1%)	557 (99.8%)	< 0.01	55 (15.5%)	138 (100.0%)	< 0.01			
Hospital LOS (days)	11.92 (6.70–22.11)	10.72 (4.59–17.97)	< 0.01	11.98 (7.07–21.92)	9.52 (5.05–17.67)	< 0.01	17.92 (11.24–29.86)	14.26 (6.93–23.90)	0.13
ICU LOS (days)	3.21 (1.92–6.46)	4.75 (2.44–9.39)	< 0.01	3.32 (1.91–6.17)	5.75 (2.78–9.71)	< 0.01	17.92 (11.24–29.86)	14.26 (6.93–23.90)	0.13

Between-cohort heterogeneity

The three cohorts differed significantly in demographics, clinical features, disease severity, and treatments ($P < 0.05$), indicating between-cohort heterogeneity (Table 3). Relative to the training and internal validation sets, external validation patients had lower body weight and BMI and lower prevalences of hypertension, diabetes, atrial fibrillation, chronic viral hepatitis, and alcoholic hepatitis, but higher SOFA scores (median 8.00, IQR 5.00–10.00 vs 5.00, IQR 3.00–7.00) and higher rates of esophagogastric varices (55.7%) and gastrointestinal bleeding (25.3%). Potassium and calcium were lower and chloride higher in the external cohort, whereas glucose was similar to the development cohorts. Regarding treatment and outcomes, CRRT (3.8%) and invasive mechanical ventilation (21.5%) were used less often in the external cohort, whereas the liver-transplantation rate was highest (17.7%) and in-hospital mortality lowest (9.3%). These findings may relate to between-center practice differences, etiology mix (chronic viral hepatitis 97.0% and alcoholic hepatitis only 11.4% in the external cohort), and the smaller sample size.

Table 3
Baseline characteristics across the training, internal validation, and external validation cohorts

Variable	Training set(n = 1970)	Internal validation(n = 493)	External validation(n = 237)	P value
Age (years)	58.00 (51.00–65.00)	58.00 (51.00–66.00)	58.00 (49.00–68.00)	0.93
Female sex	700 (35.5%)	168 (34.1%)	80 (33.8%)	0.75
Weight (kg)	82.40 (70.00–97.80)	82.00 (69.10–98.30)	62.00 (56.72–70.00)	< 0.01
BMI (kg/m²)	28.34 (24.49–33.30)	28.98 (23.96–34.41)	21.85 (20.13–24.48)	< 0.01
SOFA score	5.00 (3.00–7.00)	4.00 (3.00–7.00)	8.00 (5.00–10.00)	< 0.01
GCS score	15.00 (15.00–15.00)	15.00 (8.00–15.00)	< 0.01	
SAPS II score	41.00 (32.00–52.00)	42.00 (33.00–52.00)	37.00 (27.00–47.00)	< 0.01
LODS score	7.00 (4.00–9.00)	7.00 (5.00–9.00)	6.00 (3.00–8.00)	0.03
OASIS score	34.00 (28.00–40.00)	34.00 (28.00–41.00)	28.00 (22.00–34.00)	< 0.01
MELD score	25.06 (17.54–32.63)	25.39 (18.00–32.78)	18.54 (14.00–25.00)	< 0.01
ABIC score	8.65 (7.67–9.83)	8.67 (7.74–9.88)	8.23 (7.05–9.34)	0.00
Urine output (mL)	1128.50 (600.00–1790.00)	1175.00 (588.00–1846.99)	962.59 (507.94–1170.00)	0.00
Hypertension	718 (36.4%)	178 (36.1%)	38 (16.0%)	< 0.01
Atherosclerotic heart disease	236 (12.0%)	64 (13.0%)	20 (8.4%)	0.19
Atrial fibrillation	401 (20.4%)	105 (21.3%)	15 (6.3%)	< 0.01
Diabetes mellitus	614 (31.2%)	166 (33.7%)	47 (19.8%)	0.00
Chronic kidney disease	430 (21.8%)	105 (21.3%)	29 (12.2%)	0.00
Chronic viral hepatitis	336 (17.1%)	87 (17.6%)	230 (97.0%)	< 0.01
Alcoholic hepatitis	1073 (54.5%)	270 (54.8%)	27 (11.4%)	< 0.01
Drug-induced liver injury	12 (0.6%)	1 (0.2%)	7 (3.0%)	0.00
Cholestatic liver disease	54 (2.7%)	10 (2.0%)	25 (10.5%)	< 0.01
Autoimmune hepatitis	21 (1.1%)	4 (0.8%)	15 (6.3%)	< 0.01
Hepatic encephalopathy	201 (10.2%)	49 (9.9%)	36 (15.2%)	0.05
Esophagogastric varices	519 (26.3%)	137 (27.8%)	132 (55.7%)	< 0.01
Gastrointestinal bleeding	127 (6.4%)	35 (7.1%)	60 (25.3%)	< 0.01
Hepatic failure	421 (21.4%)	125 (25.4%)	42 (17.7%)	0.05
Hepatorenal syndrome	302 (15.3%)	73 (14.8%)	14 (5.9%)	0.00
Acute liver failure	0 (0.0%)	9 (3.8%)	< 0.01	
Peritonitis	147 (7.5%)	41 (8.3%)	22 (9.3%)	0.54
Ascites	560 (28.4%)	147 (29.8%)	98 (41.4%)	0.00
Septic shock	424 (21.5%)	99 (20.1%)	35 (14.8%)	0.05
Heart rate (beats/min)	106.00 (93.00–119.75)	105.00 (91.00–120.00)	110.00 (95.00–126.00)	0.00
SBP (mmHg)	86.00 (79.00–95.00)	85.50 (77.00–96.00)	83.00 (65.00–98.50)	0.00
DBP (mmHg)	45.00 (39.00–51.00)	44.00 (37.00–50.00)	42.00 (34.00–52.00)	0.01

Variable	Training set(n = 1970)	Internal validation(n = 493)	External validation(n = 237)	P value
MAP (mmHg)	57.00 (51.00–64.00)	56.00 (50.00–63.00)	56.00 (45.00–68.00)	0.01
Respiratory rate (breaths/min)	27.00 (23.06–32.00)	27.00 (23.00–32.00)	29.00 (25.00–30.00)	0.16
Temperature (°C)	37.22 (36.94–37.67)	37.17 (36.89–37.56)	36.91 (36.70–37.20)	< 0.01
SpO_2 (%)	93.00 (90.00–95.00)	91.00 (84.00–95.00)	< 0.01	
Glucose (mmol/L)	9.03 (7.11–12.72)	9 (7.17–12.17)	10.01 (7.12–12.30)	0.00
WBC (×10⁹/L)	12.60 (8.20–18.10)	13.10 (8.70–19.40)	11.05 (6.47–14.02)	< 0.01
Hemoglobin (g/L)	85 (72–100)	84 (72–98)	82.75 (65.00–99.00)	0.06
Hematocrit max (%)	30.60 (26.90–35.20)	30.50 (26.70–35.00)	27.60 (22.50–32.20)	< 0.01
Hematocrit min (%)	25.50 (21.70–29.80)	25.50 (21.70–29.40)	24.72 (19.80–28.70)	0.03
Platelets (×10⁹/L)	81.00 (51.00–133.00)	84.00 (51.00–131.00)	85.00 (55.00–118.00)	0.67
INR	1.90 (1.50–2.48)	1.80 (1.50–2.40)	1.65 (1.35–2.01)	0.00
PT (s)	20.10 (16.10–26.50)	19.60 (16.30–26.20)	18.00 (14.70–22.07)	0.00
PTT (s)	40.80 (33.50–56.30)	41.90 (33.70–56.90)	18.00 (14.70–22.07)	0.21
Total bilirubin (μmol/L)	54.72 (22.23–135.1)	56.43 (25.65–130.0)	43.51 (21.20–95.52)	0.67
Albumin (g/L)	28 (24–32)	28.9 (25–33)	28.50 (25.50–32.00)	0.13
ALT (U/L)	42.00 (23.25–119.00)	38.00 (22.00–83.82)	57.46 (21.00–175.00)	0.01
AST (U/L)	93.00 (47.00–262.75)	83.00 (43.00–180.00)	109.00 (42.00–318.05)	0.04
BUN (mmol/L)	11.07 (6.79–18.93)	11.79 (6.79–18.57)	9.55 (6.59–14.47)	0.17
Creatinine (μmol/L)	132.6 (79.56–229.8)	132.6 (79.56–238.7)	99.80 (66.80–175.80)	0.00
Sodium (mmol/L)	139.00 (135.00–142.00)	139.10 (136.00–143.50)	0.04	
Potassium (mmol/L)	4.50 (4.10–5.20)	4.20 (3.69–4.66)	< 0.01	
Calcium (mmol/L)	2.12 (2–2.27)	2.03 (1.91–2.20)	< 0.01	
Chloride min (mmol/L)	101.00 (95.00–105.00)	100.00 (95.00–105.00)	105.00 (101.10–109.30)	< 0.01
Chloride max (mmol/L)	105.00 (100.00–109.00)	107.00 (103.90–110.60)	< 0.01	
MCH (pg)	31.70 (29.62–34.00)	31.60 (29.40–34.00)	31.50 (30.00–33.10)	0.11
MCHC (g/L)	331.7 (321–341)	329.7 (319–340)	31.50 (30.00–33.10)	0.04
MCV (fL)	95.95 (90.00–101.75)	95.00 (90.00–101.00)	93.80 (89.20–99.50)	0.01
RDW (%)	16.60 (15.10–18.60)	16.90 (15.20–18.90)	17.53 (15.80–18.70)	< 0.01
Lactate (mmol/L)	3.25 (2.04–4.80)	3.27 (2.00–4.90)	3.90 (2.00–5.64)	0.55
CRRT	342 (17.4%)	87 (17.6%)	9 (3.8%)	< 0.01
Invasive mechanical ventilation	1020 (51.8%)	259 (52.5%)	51 (21.5%)	< 0.01
Liver transplantation	205 (10.4%)	47 (9.5%)	42 (17.7%)	0.00
Pulmonary infection	462 (23.5%)	116 (23.5%)	124 (52.3%)	< 0.01
Urinary tract infection	14 (0.7%)	9 (1.8%)	15 (6.3%)	< 0.01

Variable	Training set(n = 1970)	Internal validation(n = 493)	External validation(n = 237)	P value
Bloodstream infection	744 (37.8%)	191 (38.7%)	30 (12.7%)	< 0.01
In-hospital mortality	558 (28.3%)	138 (28.0%)	22 (9.3%)	< 0.01
28-day mortality	599 (30.4%)	150 (30.4%)		
90-day mortality	798 (40.5%)	193 (39.1%)		
Hospital LOS (days)	11.75 (6.06–21.21)	11.06 (6.39–20.85)	17.44 (11.12–29.34)	< 0.01
ICU LOS (days)	3.67 (2.01–7.28)	3.67 (2.02–7.18)	17.44 (11.12–29.34)	0.12

To explore the relationship between cirrhosis etiology and mortality, subgroup analyses by hepatitis etiology were performed in the full MIMIC cohort (n = 2463), and a forest plot of mortality risk was generated (Fig. 2). Comparing in-hospital mortality by presence or absence of each hepatitis category (nonmutually exclusive), mortality was lower among patients with chronic viral hepatitis than among those without (21.04% vs 29.75%; P < 0.01) and higher among patients with alcoholic hepatitis than among those without (30.53% vs 25.54%; P < 0.01). These differences may partly explain between-cohort mortality heterogeneity.

Machine-learning model performance

Model comparison and selection of the final model

Performance metrics for each model on the training, internal validation, and external validation sets are shown in Tables 4 and 5. Overall, XGBoost demonstrated the best and most robust predictive performance and was selected as the final model. AUROCs for XGBoost were 0.834 on the test set (cross-validation), 0.833 on internal validation, and 0.776 on external validation, yielding a three-set mean of 0.814 and relatively stable discrimination across data distributions. LightGBM and lasso ranked second and third by three-set mean AUROC (0.806 and 0.791, respectively). ROC curves are shown in Fig. 3.

Table 4
AUROC of each model across datasets

Model	Training setAUROC	Test AUROC	Internal AUROC	External AUROC
XGBoost	0.940	0.834	0.833	0.776
LightGBM	0.972	0.833	0.832	0.753
Lasso	0.861	0.838	0.821	0.714
Random Forest	0.905	0.816	0.787	0.699
Decision Tree	0.831	0.737	0.739	0.615
Ridge	0.861	0.836	0.822	0.712
Logistic	0.861	0.836	0.822	0.712

In the training set, all seven models showed good discrimination; XGBoost ranked among the top models for AUROC, accuracy, precision, recall, and F1 score and had the lowest Brier score. It achieved relatively high precision (0.870) together with substantial recall (0.672). At a probability threshold of 0.5, sensitivity, specificity, accuracy, precision, recall, F1 score, and Brier score for each model and dataset are reported in Table 5. Other models had important limitations: LightGBM performed reasonably on internal validation but had a lower external AUROC (0.753) than XGBoost; the external AUROC of lasso fell to 0.656; random forest had high precision (0.726) but very low recall (0.314), implying a high risk of missed high-risk cases; and logistic and ridge regression were not outstanding overall. Although XGBoost was not first on every individual metric, it provided the best overall balance of discrimination robustness, predictive accuracy, and clinical operability; subsequent feature-importance and SHAP analyses were based on this model.

Table 5
Detailed performance metrics at a probability threshold of 0.5

	Model	AUROC	Accuracy	Precision	Recall	Sensitivity	Specificity	F1 score	Brier score
Training set	XGBoost	0.940	0.879	0.870	0.672	0.672	0.960	0.758	0.093
	LightGBM	0.972	0.924	0.932	0.789	0.789	0.977	0.854	0.071
	Lasso	0.861	0.813	0.726	0.545	0.545	0.919	0.622	0.132
	Random Forest	0.905	0.828	0.887	0.450	0.450	0.977	0.597	0.127
	Decision Tree	0.831	0.809	0.720	0.530	0.530	0.919	0.611	0.135
	Ridge	0.861	0.814	0.723	0.557	0.557	0.916	0.630	0.132
	Logistic	0.861	0.814	0.723	0.557	0.557	0.916	0.630	0.132
Test set	XGBoost	0.834	0.793	0.687	0.496	0.496	0.911	0.576	0.142
	LightGBM	0.833	0.802	0.711	0.504	0.504	0.919	0.590	0.143
	Lasso	0.838	0.792	0.671	0.520	0.520	0.899	0.586	0.142
	Random Forest	0.816	0.786	0.788	0.333	0.333	0.965	0.469	0.154
	Decision Tree	0.737	0.740	0.558	0.398	0.398	0.875	0.464	0.183
	Ridge	0.836	0.792	0.670	0.525	0.525	0.898	0.589	0.143
	Logistic	0.836	0.792	0.670	0.525	0.525	0.898	0.589	0.143
Internal validation	XGBoost	0.833	0.805	0.706	0.522	0.522	0.915	0.600	0.142
	LightGBM	0.832	0.803	0.703	0.514	0.514	0.915	0.594	0.141
	Lasso	0.821	0.789	0.655	0.522	0.522	0.893	0.581	0.149
	Random Forest	0.787	0.783	0.746	0.341	0.341	0.955	0.468	0.158
	Decision Tree	0.739	0.751	0.581	0.391	0.391	0.890	0.468	0.176
	Ridge	0.822	0.789	0.652	0.529	0.529	0.890	0.584	0.149
	Logistic	0.822	0.789	0.652	0.529	0.529	0.890	0.584	0.149
External validation	XGBoost	0.776	0.886	0.400	0.455	0.455	0.930	0.426	0.094
	LightGBM	0.753	0.878	0.333	0.318	0.318	0.935	0.326	0.096
	Lasso	0.714	0.869	0.263	0.227	0.227	0.935	0.244	0.106
	Random Forest	0.699	0.895	0.364	0.182	0.182	0.967	0.242	0.106
	Decision Tree	0.615	0.869	0.263	0.227	0.227	0.935	0.244	0.122
	Ridge	0.712	0.861	0.238	0.227	0.227	0.926	0.233	0.107
	Logistic	0.712	0.861	0.238	0.227	0.227	0.926	0.233	0.107

Calibration curve

The XGBoost calibration curve is shown in Fig. 4. Agreement with the ideal line was closer at lower predicted probabilities; at higher predicted probabilities (P ≥ 0.6), the observed positive rate rose more steeply with predicted probability, while overall calibration remained acceptable. Together with an internal validation AUROC of 0.833, these findings suggest that the model can discriminate high- from low-risk patients while providing reasonably credible absolute risk estimates.

Comparison of the final model with conventional scores

On the test and internal validation sets, XGBoost showed higher discrimination than ABIC, GCS, LODS, MELD, OASIS, SAPS II, and SOFA (Fig. 5). AUROCs were 0.834 and 0.833, respectively, exceeding those of the conventional scores; Brier scores were 0.142 and 0.142, and F1 scores also ranked among the best. On external validation, precision, recall, and F1 for several conventional scores fell to 0 or near 0, whereas XGBoost retained the highest AUROC, an accuracy of 0.886, and acceptable calibration.

Model interpretability (SHAP)

SHAP was used to examine global feature importance, direction of feature effects, and individual predictions. The five features with the largest contributions were liver transplantation status, total bilirubin (TBIL), prothrombin time (PT), bloodstream infection, and urine output (Fig. 6), indicating that model decisions were closely related to coagulation, transplant-related status, infection severity, liver function, and basic vital signs.

The SHAP summary plot (Fig. 7) shows the direction and nonlinearity of feature effects on model output. Liver transplantation was associated with lower predicted mortality risk. Higher total bilirubin and prolonged PT generally contributed to higher mortality risk; bloodstream infection and reduced urine output were likewise usually associated with increased risk. Features such as PT may show nonlinear or bidirectional patterns and should be interpreted with summary and dependence plots rather than by simple linear extrapolation.

In a representative high-risk SHAP force plot (Fig. 8), the predicted risk of death was driven mainly by high bilirubin, worsening coagulopathy, bloodstream infection, and mechanical ventilation, illustrating individualized risk assessment based on multiple features.

Discussion

Using two independent cohorts from MIMIC-IV and Zhongnan Hospital of Wuhan University, we developed and validated a model for in-hospital mortality after ICU admission in patients with cirrhosis and sepsis. Among logistic regression, ridge, lasso, decision tree, random forest, XGBoost, and LightGBM, XGBoost performed best, with AUROCs of 0.834 (test set), 0.833 (internal validation), and 0.776 (external validation) and a three-set mean of 0.814. LightGBM and lasso ranked second and third (three-set means 0.806 and 0.791), indicating that both gradient-boosted trees and regularized linear models had useful discrimination in this population, whereas XGBoost was more balanced across the three datasets. At a threshold of 0.5, external validation sensitivity, specificity, and accuracy were 0.455, 0.930, and 0.886, respectively; the threshold may be adjusted according to clinical tolerance for missed and false-positive classifications.

Compared with SOFA, SAPS II, OASIS, LODS, MELD, and ABIC, the present model had higher AUROCs on the test and internal validation sets and also superior discrimination on independent external validation. Conventional scores are typically single-time-point linear or weighted sums and capture nonlinearities and interactions among organ function, laboratory measures, and treatments only to a limited extent; XGBoost can learn such relationships through tree ensembles and achieved better prediction under the features and sample size of this study[23, 29].

SHAP analysis showed that variables with the highest mean absolute SHAP values included liver transplantation, organ-function–related measures, and laboratory indices (Figs. 6 and 7), consistent with clinical understanding of prognosis in cirrhosis with infection. Summary and dependence plots further illustrated the direction and magnitude of feature effects, and the individual force plot (Fig. 8) clarified case-level prediction rationale. These findings suggest that an XGBoost model using first-day ICU information outperforms conventional scores, offers useful interpretability, and may serve as an adjunct for early risk stratification.

Comparison with previous studies

When ICU patients with cirrhosis develop sepsis, systemic inflammation and organ failure can substantially increase in-hospital mortality; early risk assessment may help allocate monitoring resources, initiate targeted antimicrobial and organ support, and support prognostic communication with patients and families. Prior work has largely relied on conventional scores such as SOFA, MELD, and Child–Pugh[23, 29, 34]. These scores are simple and widely used but are mostly linear or fixed-weight sums, capture multivariable nonlinearities and interactions only limitedly, and often derive variables and weights from expert opinion or single-center data, so performance may vary across populations and care settings. More recently, the LIVERAID-ICU score was developed specifically for ICU patients with cirrhosis and infection and showed good discrimination using a small set of bedside variables[35]; however, it remains a conventional additive score rather than a multi-algorithm machine-learning.

Machine-learning prognostic models based on large critical-care databases such as MIMIC and eICU have increased in recent years and have shown discrimination superior to conventional scores in settings such as sepsis and acute kidney injury[25, 36]. Models for cirrhosis or liver failure have also been reported, but many focus on ACLF, transplant candidates, or general hospitalized cirrhosis[24]. More recently, several groups have specifically targeted cirrhosis with sepsis: Lin and colleagues developed a MIMIC-IV–based nomogram that outperformed SOFA, MELD-Na, and ABIC[30]; Hu et al. added external validation using eICU data[31]; Xu et al. reported a multicentre ICU model with temporal and external validation[32]; and an interpretable early 28-day mortality model has also been proposed[33]. Relative to these studies, the present work focuses on an ICU-restricted cirrhosis–sepsis cohort, systematically compares seven algorithms (including XGBoost and LightGBM) against a panel of conventional scores[23, 29], and includes external validation in an independent Chinese tertiary-hospital cohort. Although external AUROC declined to 0.776, it remained higher than that of conventional scores. Such attenuation on external validation is common and may reflect differences in case mix, practice patterns, or data-collection procedures[37]. SHAP-based feature importance and case-level explanations may also facilitate clinical understanding and acceptance[27].

Strengths

First, the model was based on machine-learning algorithms with a relatively large development sample (training n = 1970; internal validation n = 493), which favors learning stable predictive patterns; external validation used Zhongnan Hospital data from 2015–2025 (n = 237), independent of the development cohort in geography, calendar time, and health-care system, allowing preliminary assessment of generalizability to heterogeneous data, with performance superior to conventional scores. Second, predictors were limited to variables available on ICU day 1, supporting early risk assessment consistent with the need for prompt identification of high-risk patients. Third, SHAP provided global importance, summary plots, dependence plots, and individual explanations, improving interpretability of the tree model. Despite substantial between-cohort heterogeneity, external validation performance remained useful, suggesting that the model may capture core drivers of mortality risk beyond incidental features of a single population.

Limitations

First, this was a retrospective study using existing electronic health records and MIMIC-IV; variable definitions, recording times, and missingness patterns were influenced by local practice, and prospective validation with prespecified variables—as encouraged by TRIPOD + AI[26]—was not performed. Second, at a threshold of 0.5, XGBoost showed external sensitivity of 0.455 and specificity of 0.930, indicating that although high-risk patients can be identified early with high specificity, sensitivity for patients who later deteriorate and die is limited; this relates to restriction of predictors to the first 24 ICU hours, and future work may incorporate dynamic variables. Third, population characteristics differed between MIMIC and Zhongnan Hospital. Mortality analyses suggested that in-hospital death was associated with cirrhosis etiology, which varies by country and region; although external performance was acceptable, larger multicenter domestic cohorts will be needed to develop models better tailored to local etiology, consistent with guidance on transportability of prognostic models[37]. Fourth, cause-specific deaths were not distinguished; the model can provide risk signals and explanatory factors but cannot by itself yield cause-specific treatment recommendations.

Conclusions

Among ICU patients with cirrhosis and sepsis, an XGBoost in-hospital mortality model based on first-day clinical and laboratory variables showed good discrimination on the MIMIC-IV test set, internal validation set, and Zhongnan Hospital external validation set (cross-validation AUROC 0.834, internal AUROC 0.833, external AUROC 0.776), outperformed conventional scores including SAPS II, OASIS, LODS, MELD, ABIC, and GCS, and offered useful interpretability. Future work may include prospective multicenter

validation, integration with electronic health records or clinical decision support systems, updating predictions with dynamic variables, and extension to 28-day or 90-day mortality or multiple endpoints.

Abbreviations

AUROC
area under the receiver operating characteristic curve
ICU
intensive care unit
MIMIC-IV
Medical Information Mart for Intensive Care IV
SHAP
SHapley Additive exPlanations
SOFA
Sequential Organ Failure Assessment
SAPS II
Simplified Acute Physiology Score II
OASIS
Oxford Acute Severity of Illness Score
LODS
Logistic Organ Dysfunction System
MELD
Model for End-Stage Liver Disease
ABIC
Age–Bilirubin–INR–Creatinine score
GCS
Glasgow Coma Scale.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional ethics committee of Zhongnan Hospital of Wuhan University (Approval No. 2026049k). MIMIC-IV access was obtained after completion of the required training (CITI certification number 57917614). All data were de-identified, and the requirement for informed consent was waived.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Author Contribution

ZJH, CH, and PZY conceived and designed the study. CH performed the statistical analysis. CH drafted the manuscript. YC, ZJH, MPGZ, and DYP collected the clinical data. WZX critically revised the manuscript for important intellectual content. PZY supervised

the study. All authors read and approved the final manuscript and agree to be accountable for all aspects

Acknowledgements

Not applicable

References

1. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. J Hepatol. 2019;70(1):151–71. https://doi.org/10.1016/j.jhep.2018.09.014.
2. Israelsen M, Francque S, Tsochatzis EA, Krag A. Steatotic liver disease. Lancet. 2024;404(10464):1761–78. https://doi.org/10.1016/S0140-6736(24)01811-7.
3. Devarbhavi H, Asrani SK, Arab JP, Nartey YA, Pose E, Kamath PS. Global burden of liver disease: 2023 update. J Hepatol. 2023;79(2):516–37. https://doi.org/10.1016/j.jhep.2023.03.017.
4. Tham EKJ, Tan DJH, Danpanichkul P, et al. The global burden of cirrhosis and other chronic liver diseases in 2021. Liver Int. 2025;45(3):e70001. https://doi.org/10.1111/liv.70001.
5. Con D, Kemp W, Majumdar A, Pilcher D, Roberts SK, Majeed A. Long-term outcomes after ICU admission in critically ill patients with liver cirrhosis: an Australian state-wide cohort study. Hepatol Commun. 2025;9(8):e0762. https://doi.org/10.1097/HC9.0000000000000762.
6. Kosuta I, Premkumar M, Reddy KR. Review article: evaluation and care of the critically ill patient with cirrhosis. Aliment Pharmacol Ther. 2024;59(12):1489–509. https://doi.org/10.1111/apt.18016.
7. Cao Z, Wong F, Choudhury AK, et al. Global prevalence and characteristics of infections and clinical outcomes in hospitalised patients with cirrhosis: a prospective cohort study for the CLEARED Consortium. Lancet Gastroenterol Hepatol. 2024;9(11):997–1009. https://doi.org/10.1016/S2468-1253(24)00224-3.
8. Trebicka J, Fernandez J, Papp M, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. J Hepatol. 2020;73(4):842–54. https://doi.org/10.1016/j.jhep.2020.06.013.
9. Arvaniti V, D'Amico G, Fede G, et al. Infections in patients with cirrhosis increase mortality four-fold and should be used in determining prognosis. Gastroenterology. 2010;139(4):1246–56. https://doi.org/10.1053/j.gastro.2010.06.019. 1256.e1–5.
10. Bajaj JS, Kamath PS, Reddy KR. The evolving challenge of infections in cirrhosis. N Engl J Med. 2021;384(24):2317–30. https://doi.org/10.1056/NEJMra2021808.
11. Piano S, Bunchorntavakul C, Marciano S, Reddy KR. Infections in cirrhosis. Lancet Gastroenterol Hepatol. 2024;9(8):745–57. https://doi.org/10.1016/S2468-1253(24)00078-5.
12. Albillos A, Martin-Mateos R, Van der Merwe S, Wiest R, Jalan R, Álvarez-Mon M. Cirrhosis-associated immune dysfunction. Nat Rev Gastroenterol Hepatol. 2022;19(2):112–34. https://doi.org/10.1038/s41575-021-00520-7.
13. Arroyo V, Angeli P, Moreau R, et al. The systemic inflammation hypothesis: Towards a new paradigm of acute decompensation and multiorgan failure in cirrhosis. J Hepatol. 2021;74(3):670–85. https://doi.org/10.1016/j.jhep.2020.11.048.
14. Piano S, Singh V, Caraceni P, et al. Epidemiology and effects of bacterial infections in patients with cirrhosis worldwide. Gastroenterology. 2019;156(5):1368–e8010. https://doi.org/10.1053/j.gastro.2018.12.005.
15. Prin M, Bakker J, Wagener G. Hepatosplanchnic circulation in cirrhosis and sepsis. World J Gastroenterol. 2015;21(9):2582–92. https://doi.org/10.3748/wjg.v21.i9.2582.
16. Cai J, Yip TC, Wong GL, et al. Global epidemiology of bacterial infections in patients with cirrhosis: a systematic review and meta-analysis. J Hepatol. 2026;84(6):1089–101. https://doi.org/10.1016/j.jhep.2026.02.007.
17. Strnad P, Tacke F, Koch A, Trautwein C. Liver — guardian, modifier and target of sepsis. Nat Rev Gastroenterol Hepatol. 2017;14(1):55–66. https://doi.org/10.1038/nrgastro.2016.168.
18. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801–10. https://doi.org/10.1001/jama.2016.0287.
19. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med. 2021;47(11):1181–247. https://doi.org/10.1007/s00134-021-06506-y.

20. Jung RG, Gupta A, Stotts C, et al. Prognostic factors associated with mortality in septic shock: a systematic review and meta-analysis. Lancet Respir Med. 2026;14(3):225–32. https://doi.org/10.1016/S2213-2600(25)00397-2.
21. Fleischmann-Struzek C, Mellhammar L, Rose N, et al. Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis. Intensive Care Med. 2020;46(8):1552–62. https://doi.org/10.1007/s00134-020-06151-x.
22. Girish V, Maiwall R. Revisiting septic shock in cirrhosis: a call for personalized management. Expert Rev Gastroenterol Hepatol. 2024;18(12):795–813. https://doi.org/10.1080/17474124.2024.2443813.
23. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA. 2001;286(14):1754–8. https://doi.org/10.1001/jama.286.14.1754.
24. Wang ZJ, Li FY, Cai JJ, Xue ZT, Zhou Y, Wang Z. Construction and validation of a machine learning-based prediction model for short-term mortality in critically ill patients with liver cirrhosis. Clin Res Hepatol Gastroenterol. 2025;49(1):102507. https://doi.org/10.1016/j.clinre.2024.102507.
25. Rahman MS, Islam KR, Prithula J, et al. Machine learning-based prognostic model for 30-day mortality prediction in Sepsis-3. BMC Med Inf Decis Mak. 2024;24(1):249. https://doi.org/10.1186/s12911-024-02655-4.
26. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74. https://doi.org/10.48550/arXiv.1705.07874.
28. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x.
29. Wiesner R, Edwards E, Freeman R, et al. Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology. 2003;124(1):91–6. https://doi.org/10.1053/gast.2003.50016.
30. Lin HR, Liao QX, Lin XX, Zhou Y, Lin JD, Xiao XJ. Development of a nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis. Sci Rep. 2024;14(1):9759. https://doi.org/10.1038/s41598-024-60305-1.
31. Hu Y, Zhang L, Yin J. Development and external validation of a predictive model for in-hospital mortality in patients with liver cirrhosis and sepsis. Sci Rep. 2026;16:15885. https://doi.org/10.1038/s41598-026-43991-x.
32. Xu Z, Yang Q, Peng D, et al. Development and validation of a prediction model for in-hospital mortality in intensive care unit patients with cirrhosis and sepsis: a multicentre retrospective cohort study. Front Med (Lausanne). 2026;13:1759988. https://doi.org/10.3389/fmed.2026.1759988.
33. Xu X, Li J, Yu H, Huang J. Interpretable model for early prediction of 28-day mortality in patients with cirrhosis and sepsis: a multi-cohort ICU study. BMC Gastroenterol. 2026;26:123. https://doi.org/10.1186/s12876-026-04620-z.
34. Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. Br J Surg. 1973;60(8):646–9. https://doi.org/10.1002/bjs.1800600817.
35. Hoppmann H, Zeman F, Stöckert P, et al. The LIVERAID (LIVER And Infectious Diseases)-ICU score predicts in-hospital mortality in liver cirrhosis patients with infections in the intensive care unit. BMJ Open Gastroenterol. 2024;11(1):e001482. https://doi.org/10.1136/bmjgast-2024-001482.
36. Hou N, Li M, He L, et al. Predicting 30-day mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. J Transl Med. 2020;18(1):462. https://doi.org/10.1186/s12967-020-02620-5.
37. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49–58. https://doi.org/10.1093/ckj/sfaa188.

Figures

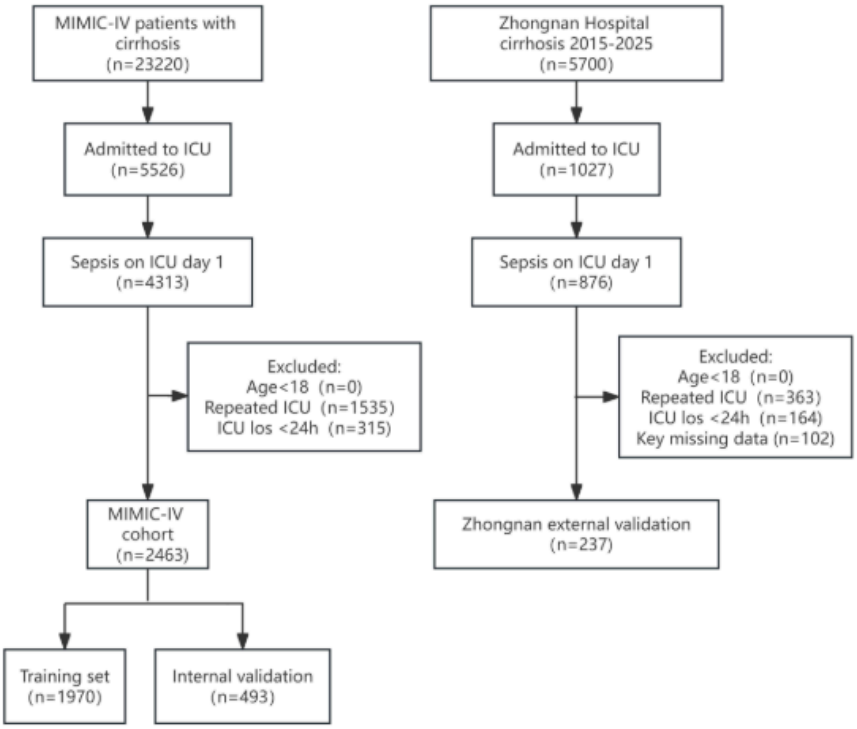


Figure 1

Study flowchart.

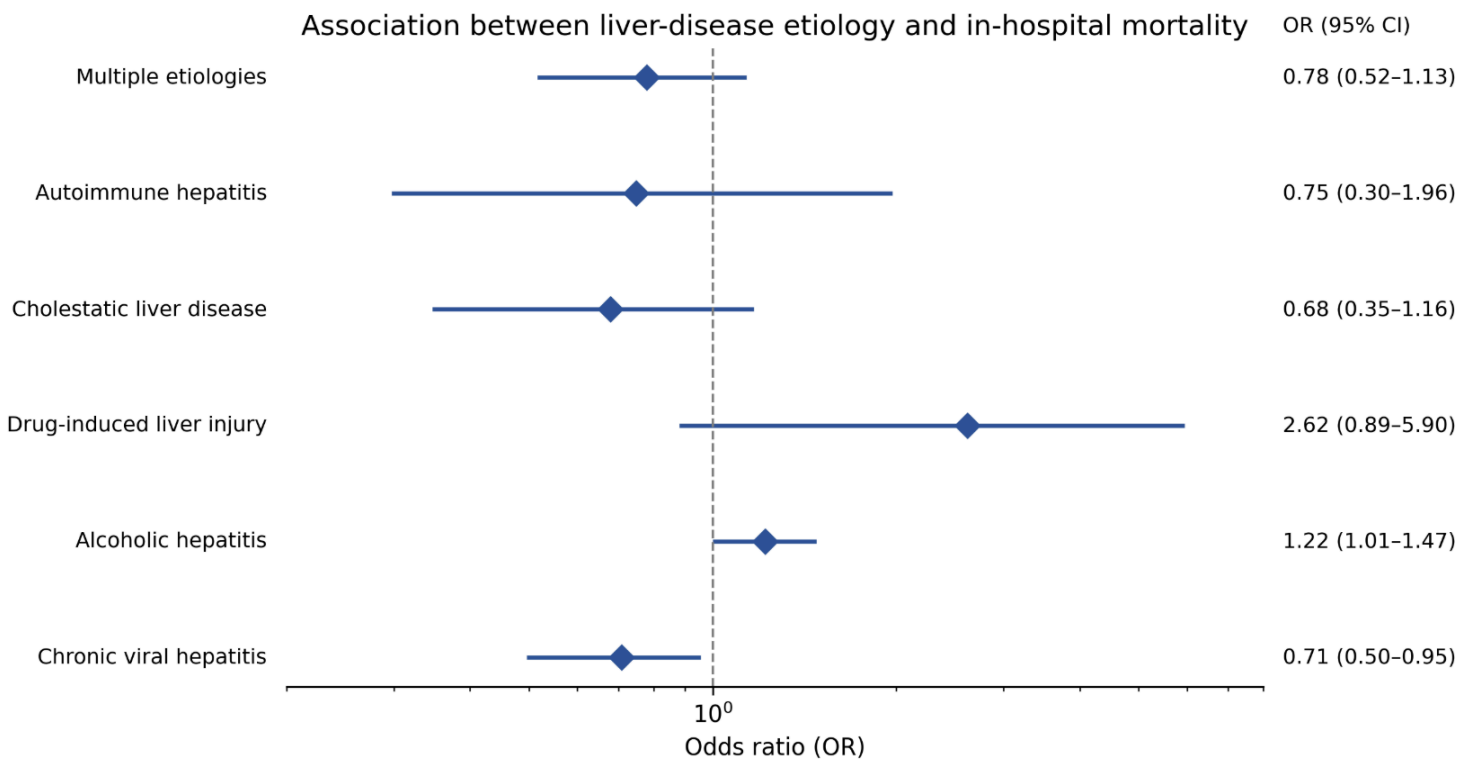


Figure 2

Forest plot of associations between liver-disease etiology and in-hospital mortality in the MIMIC-IV cohort.

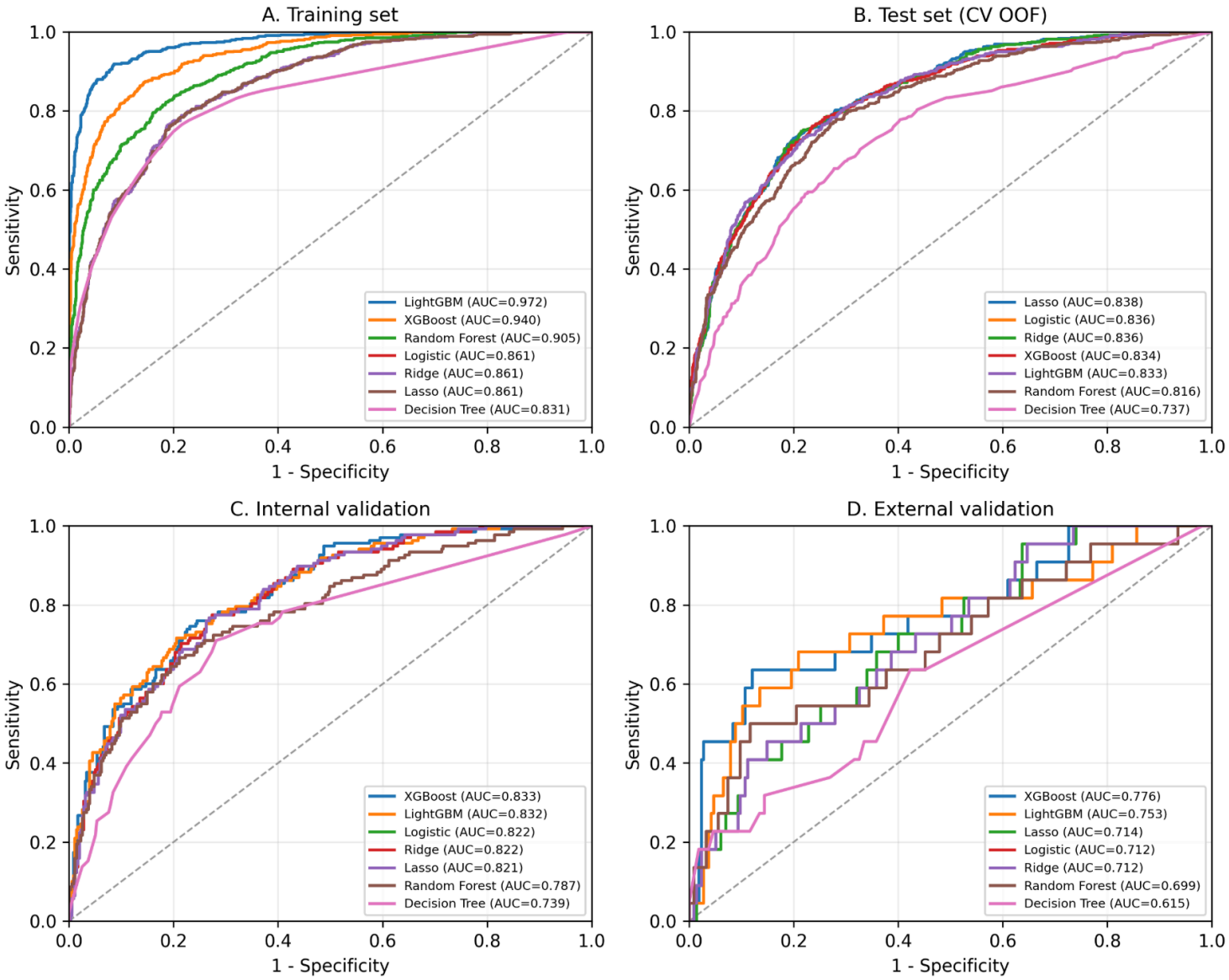


Figure 3

ROC curves of seven machine-learning models across datasets.

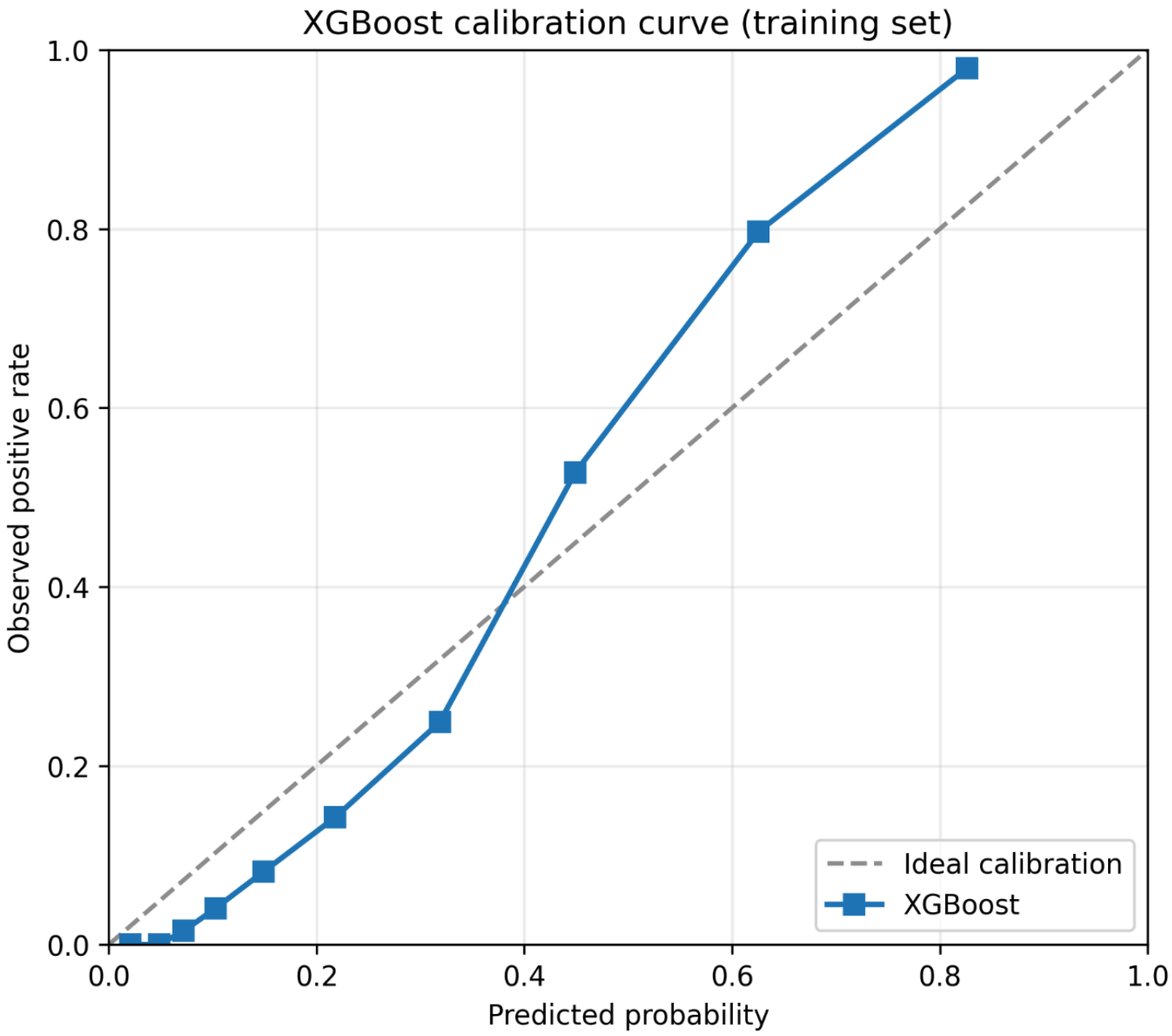


Figure 4

Calibration curve of the XGBoost model.

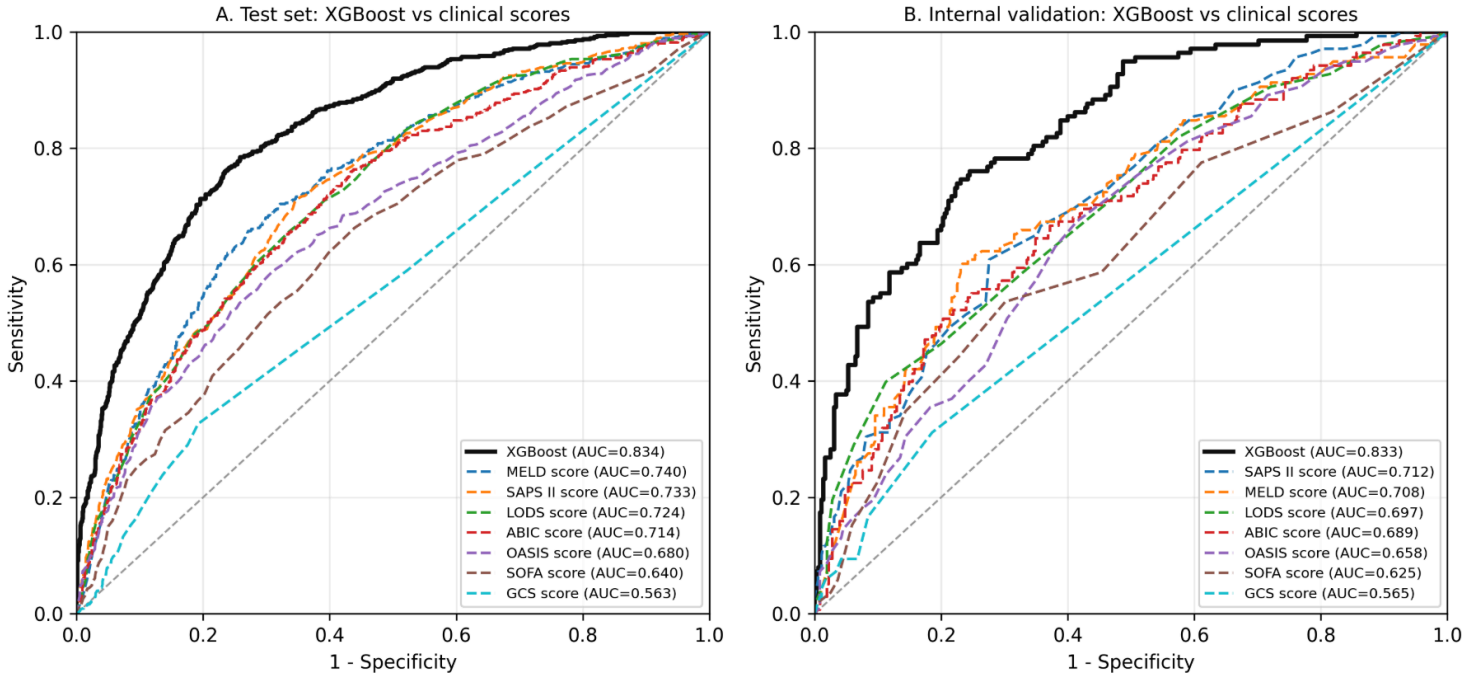


Figure 5

ROC comparison of XGBoost with conventional clinical scores.

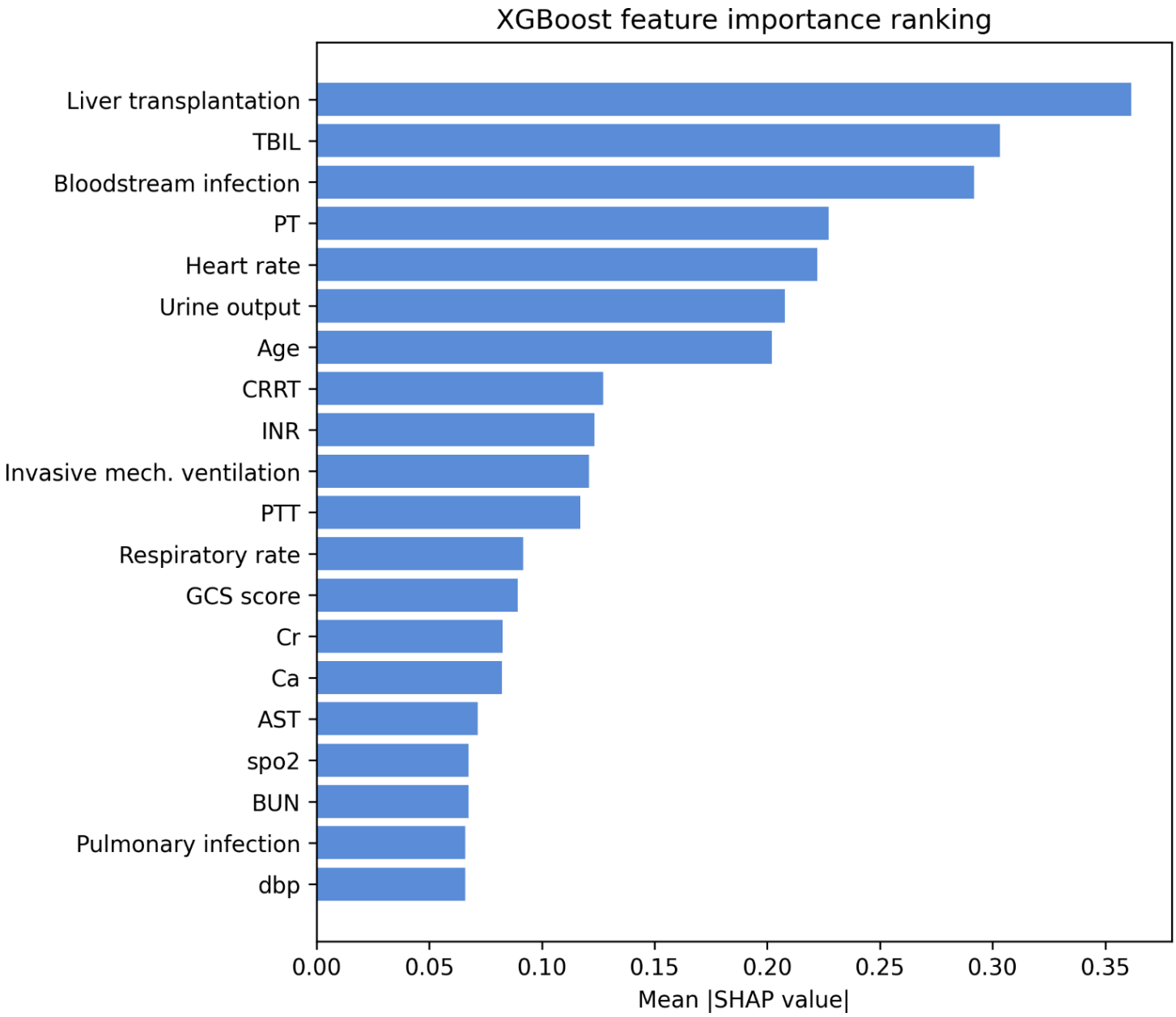


Figure 6

XGBoost feature importance ranking based on mean |SHAP| values.

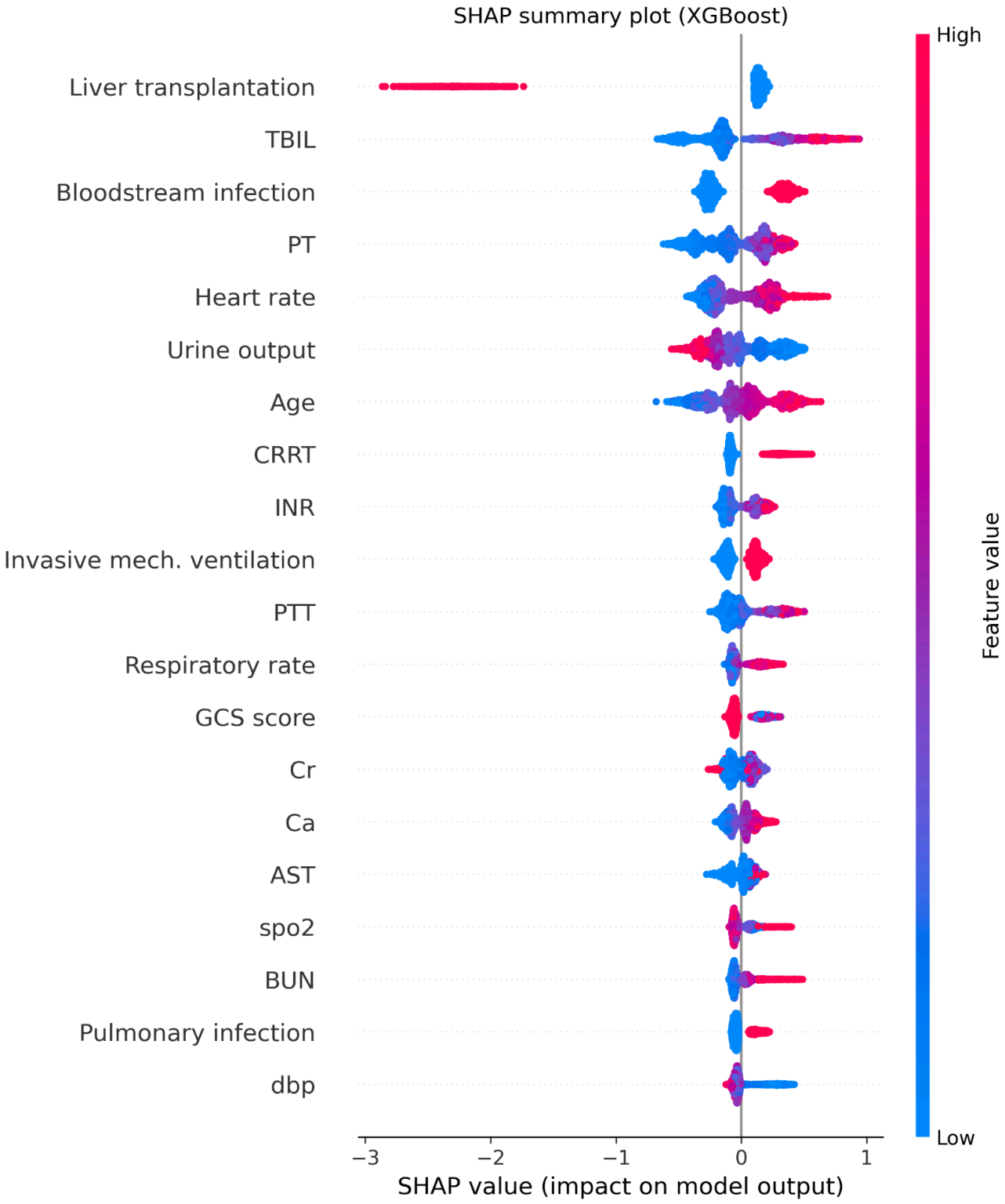


Figure 7

SHAP summary plot of the XGBoost model.

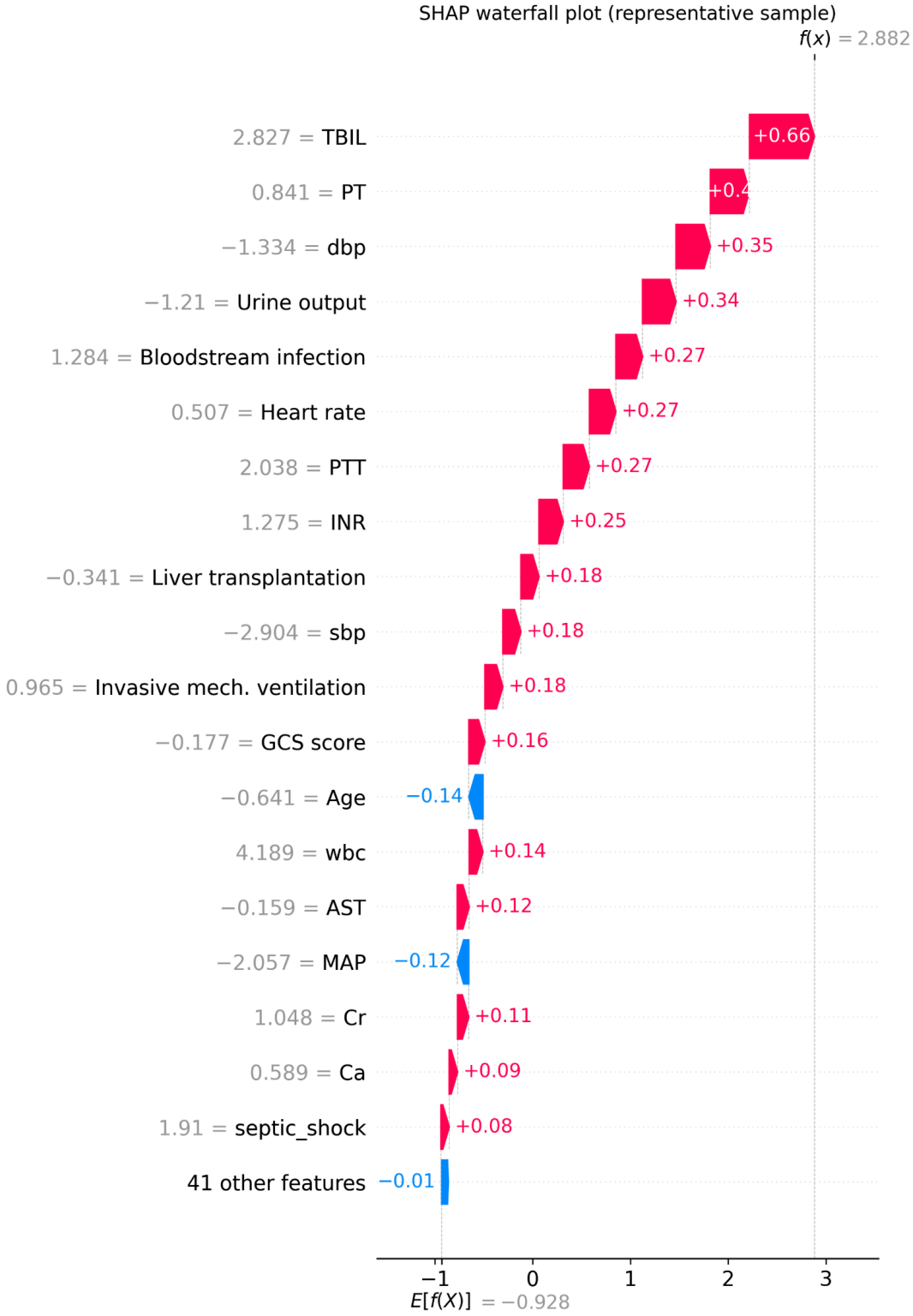


Figure 8

SHAP waterfall plot for a representative sample.