BioMedInformatics

MDPI

Article

Development of Machine Learning Models to Predict 28-Day Mortality in Patients with Sepsis-Associated Liver Injury

Yupeng Li [1], Junyi Fan [1], Kamiar Alaei [2] and Maryam Pishgar [1,*]

[1] Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA; yupengli@usc.edu (Y.L.); junyifan@usc.edu (J.F.)
[2] Department of Health Science, California State University, Long Beach, CA 90840, USA; kamiar.alaei@csulb.edu
* Correspondence: pishgar@usc.edu

Abstract

Background: Sepsis-associated liver injury (SALI) is a serious complication of sepsis that increases the risk of organ dysfunction and mortality; however, early identification of high-risk patients remains difficult due to nonspecific clinical features and complex pathophysiology. This study aimed to develop machine learning (ML) models to predict 28-day mortality in SALI patients within the first 24 h of intensive care unit (ICU) admission. **Methods:** A total of 1157 patients were included, comprising 826 from the MIMIC-IV (v2.2) database, 225 from MIMIC-III (v1.4), and 106 from eICU (v2.0). Data from MIMIC-IV were split into training and internal validation sets (7:3), while MIMIC-III and eICU served as external validation cohorts. Thirty clinically relevant features were selected. Eight ML models were evaluated using AUROC, accuracy, precision, recall, F1-score, and specificity. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) enhanced interpretability. **Results:** XGBoost model achieved the best performance, with an AUROC of 0.8556 (95% CI: 0.807–0.898), accuracy of 0.7702, recall of 0.8469, and specificity of 0.7200. SHAP identified lactate, blood urea nitrogen, heart rate, hemoglobin, and diastolic blood pressure as key predictors, while LIME provided patient-level interpretability. **Conclusions:** The XGBoost-based model may facilitate early mortality risk stratification and support clinical decision-making for SALI patients in ICU settings.

Keywords: sepsis; machine learning; sepsis-associated liver injury; XGBoost; MIMIC-IV; intensive care unit (ICU); mortality prediction; external validation

1. Introduction

Sepsis is defined by the third international consensus (Sepsis-3) as organ dysfunction resulting from a dysregulated host response to infection. It disrupts the body's finely tuned balance between inflammatory and anti-inflammatory responses, and the resulting inflammation can lead to exacerbated tissue damage, ultimately causing multiple organ dysfunction syndrome (MODS). Both innate and adaptive immune responses play a crucial role in the pathogenesis of sepsis [1,2]. According to data from the Global Burden of Disease (2020), there were approximately 48.9 million cases of sepsis worldwide in 2017, resulting in 11 million deaths, which accounted for 19.7% of all global deaths [3]. Infants, children, and the elderly (over 65 years old), who have weaker immune systems, are particularly susceptible to sepsis and its associated mortality [4].

Sepsis patients often experience various long-term complications, with liver damage being one of the most severe. The liver is the largest solid organ in the human body,

Check for updates

Academic Editor: Carson K. Leung

Received: 20 November 2025
Revised: 26 December 2025
Accepted: 12 January 2026
Published: 13 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

performing essential functions such as metabolism, detoxification, and bile secretion, while also playing a vital role in immune defense by eliminating pathogens and toxins and regulating inflammatory responses [5,6]. Consequently, liver dysfunction disrupts the body's metabolic and immune homeostasis, significantly increasing the risk of infection spread and multi-organ failure [7]. Sepsis-associated liver injury (SALI) is a common and serious complication of sepsis, occurring in approximately 34% to 46% of patients [8]. SALI is typically classified into three types: ischemic-hypoxic liver injury, cholestatic liver injury, and hepatocellular injury [9]. Due to the lack of a standardized diagnostic criterion, current research often relies on the Surviving Sepsis Campaign (SSC) Guidelines, which define SALI as total bilirubin > 2 mg/dL and INR > 1.5 [10]. Early identification and intervention for SALI are crucial, as they can guide resource allocation, optimize treatment strategies, and potentially reduce mortality.

However, the nonspecific clinical symptoms and complex etiology of sepsis in its early stages make timely diagnosis difficult. Traditional diagnostic methods are often inadequate, highlighting the need for more advanced approaches. Machine learning offers a promising alternative for predicting outcomes in critically ill patients, surpassing traditional statistical methods [11]. However, machine learning models continue to face challenges regarding generalizability across different datasets and interpretability for clinical applications.

This study aims to develop and validate an interpretable machine learning model for the early prediction of 28-day mortality in patients with sepsis-associated liver injury. By integrating multi-center data integration, comprehensive data preprocessing, feature selection, multi-model development, and dual-level interpretability analyses (SHAP and LIME), our approach improves model performance, ensures model generalizability across diverse ICU populations, and improves clinical interpretability:

Multi-center data integration: The clinical data were extracted from three large ICU datasets(MIMIC-IV, MIMIC-III, and eICU-CRD), enhancing model robustness and generalizability through multi-center external validation and ensuring stable performance across diverse patient populations.

Model interpretability and clinical insight: SHAP and LIME analyses were employed to provide transparency at both global and individual patient levels to support clinical decision-making. The analyses identified five key predictors, including Lactate Min, BUN Min, Heart Rate Max, Hemoglobin Max, and DiaBP Min, closely aligned with the established clinical pathophysiology of SALI.

Feature selection procedure: A systematic feature-selection strategy was implemented, combining statistical testing (Variance Inflation Factor), clinical expertise, and model-based importance metrics (SHAP values and permutation importance). By incorporating both minimum and maximum physiological values to capture dynamic variability, the process yielded a refined and clinically meaningful set of 30 predictors.

Model development and optimization: Eight machine learning models were evaluated, which included logistic regression, elastic-net penalized logistic regression (cv-glmnet), random forest, support vector machine (SVM), extreme gradient boosting (XGBoost), k-nearest neighbors (KNN), decision tree, and linear discriminant analysis (LDA). The XGBoost algorithm was optimized through grid search and 5-fold cross-validation. Its performance surpassed existing models (AUROC = 0.8556), achieving balanced sensitivity and specificity for reliable mortality risk prediction.

2. Materials and Methods

2.1. Data Source

This study extracted data from three large, publicly available, and comprehensive clinical care databases, the Medical Information Mart for Intensive Care IV (MIMIC-IV,

v2.2, 2008–2019) [12], the Medical Information Mart for Intensive Care III (MIMIC-III, v1.4, 2001–2012) [13], and the multi-center eICU Collaborative Research Database (eICU-CRD, v2.0, 2014–2015) [14]. These databases contain detailed patient information, such as demographics, vital signs, laboratory results, clinical notes, and survival outcomes. The MIMIC databases were developed through a collaboration between the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology and Beth Israel Deaconess Medical Center (BIDMC), with funding support from the National Institutes of Health (NIH). The eICU-CRD was created through a partnership between Philips Healthcare and the MIT Laboratory for Computational Physiology.

2.2. Study Population

The study cohort consisted of adult patients (Age $\geq$ 18) diagnosed with sepsis-associated liver injury (SALI) who had an intensive care unit (ICU) stay of at least 24 h in the MIMIC-IV database. Sepsis was defined based on the Sepsis-3 criteria [1] as suspected infections with a Sequential Organ Failure Assessment (SOFA) score $\geq$ 2. Sepsis-associated liver injury (SALI) is characterized by a total bilirubin (TBIL) level > 2 mg/dL and an international normalized ratio (INR) > 1.5, according to the Surviving Sepsis Campaign (SSC) International Guidelines [10]. Patients were excluded based on the following criteria: (1) human immunodeficiency virus (HIV) infection, (2) pregnancy, (3) absence of liver injury, (4) other etiologies of liver disease, and (5) without biochemical and coagulation tests within 24 h of ICU admission. The same inclusion and exclusion criteria were applied to the MIMIC-III database and the eICU Collaborative Research Database for external validation. In total, 826 patients from MIMIC-IV, 225 patients from MIMIC-III, and 106 patients from eICU-CRD were included in the final study cohort. The detailed patient selection process is illustrated in Figure 1.

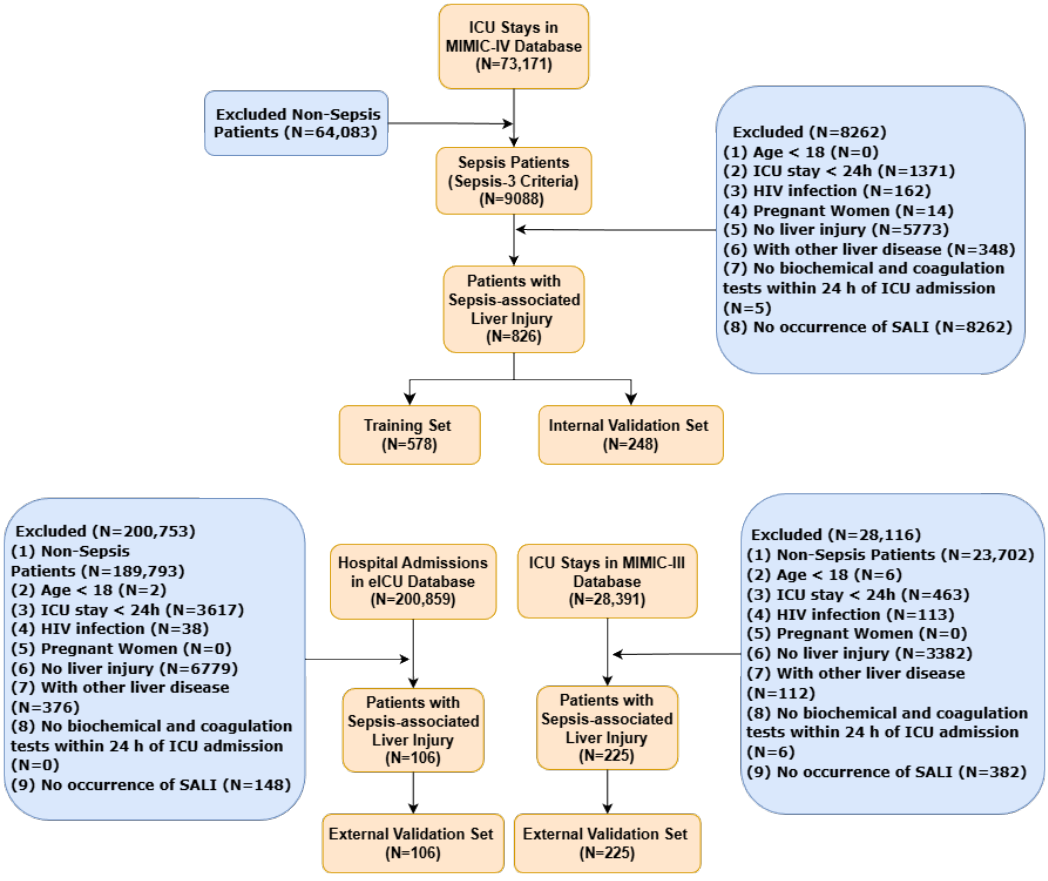


Figure 1. Patient Selection Process.

2.3. Data Collection

The data were extracted by Structured Query Language (SQL), MySQL (Version 8.0; Oracle Corporation, Austin, TX, USA), and Python (Version 3.10.12), based on existing literature [15], clinical relevance, and expert recommendations to capture early clinical status. All diagnoses were identified based on the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) and Ninth Revision, Clinical Modification (ICD-9-CM). Initially, a total of 68 clinical features were extracted from the MIMIC-IV database, categorized into 8 groups: (1) demographics (age, ICU length of stay, gender, Body Mass Index (BMI), race, admission type), (2) vital signs (heart Rate, respiratory rate, diastolic blood pressure (DBP), systolic blood pressure (SBP), temperature, oxygen saturation (SpO_2), mean arterial pressure (MAP)), (3) laboratory indicators (platelet, prothrombin time (PT), partial thromboplastin time (PTT), total bilirubin, direct bilirubin, lactate dehydrogenase, serum albumin, blood urea nitrogen (BUN), anion gap, base excess, alanine aminotransferase (ALT), alkaline phosphatase (ALP), hematocrit, serum chloride, aspartate aminotransferase (AST), creatinine, blood glucose, hemoglobin, white blood cell, PaO_2/FiO_2 ratio, red blood cell, serum sodium, bicarbonate, serum potassium, serum calcium, lactate, neutrophil-to-lymphocyte ratio (NLR), C-reactive protein (CRP), international normalized ratio (INR), magnesium, fibrinogen), (4) comorbidities (hypertension, congestive heart failure, myocardial infarction, liver disease, cerebrovascular disease, chronic pulmonary disease, renal disease, diabetes), (5) infection sites (intestinal, urinary, pulmonary, catheter-related, skin and soft tissue, abdominal cavity, central nervous system), (6) interventions (vasopressor use, mechanical ventilation, blood transfusion), (7) clinical measurements (total urine output on day 1, glomerular filtration rate (GFR)), and (8) severity scores (GCS score, SOFA score, qSOFA score, APACHE II score). The Glasgow Coma Scale (GCS) score used in this study refers to the total score, including eye opening (E), verbal response (V), and motor response (M), to evaluate the level of consciousness [16]. For vital signs and laboratory indicators, both minimum and maximum values were calculated to capture the full spectrum of dynamic physiological fluctuations during the first 24 h of ICU admission. Hemodynamic and biochemical parameters in critically ill patients often change rapidly rather than remaining in a single stable state [17]. While this approach increased feature dimensionality, multicollinearity control and the regularization mechanisms within the XGBoost framework helped mitigate the risk of overfitting. According to the Surviving Sepsis Campaign (SSC) International Guidelines [10], this study focused on 28-day mortality as a primary outcome for evaluating short-term survival in patients with sepsis. To enhance the clinical applicability of the model, the data collection timeframe was strictly defined, ensuring that all features were obtained within the first 24 h of ICU admission. The same set of features was retrieved from the MIMIC-III and eICU-CRD databases to ensure consistency between study cohorts and to minimize potential bias related to feature selection. The complete list of 68 clinical features is provided in Table 1.

2.4. Data Preprocessing

Data preprocessing is a fundamental step in machine learning and clinical research, ensuring feature comparability and robust model performance. Prior to subsequent analyses, MIMIC-IV data were randomly divided into a training set (N = 578) and an internal validation set (N = 248) at a 7:3 ratio using a stratified train-test split to avoid data leakage [18]. The outcome distribution demonstrated a relatively balanced class ratio (61:39); thus, resampling techniques were not employed. The initial dataset consisted of 106 extracted features, of which 88 features remained after excluding features with more than 20% missing values. Remaining missing values were imputed using the median, while outliers were addressed through winsorization to cap extreme values outside the 1st–99th percentile

range at the corresponding thresholds. For highly skewed variables, a log transformation was applied to improve distributional properties and stabilize variance. Multicollinearity was assessed using the variance inflation factor (VIF), and features with VIF values greater than 20 were excluded to reduce collinearity while retaining clinically relevant predictors [19,20]. After VIF-based filtering, 86 features were retained for further analysis. Additionally, redundant predictors were identified by examining pairwise correlation coefficients, and those exceeding 0.85 were removed to improve computational efficiency and reduce overfitting [21]. This correlation-based filtering step further reduced the feature set to 82 features. Finally, all numerical features were preprocessed with a standard scaler fitted on the training set and applied to the validation set. Categorical variables, such as gender and race, were encoded using label encoding and one-hot encoding to convert them into numerical formats, which facilitated compatibility with machine learning algorithms while maintaining interpretability.

Table 1. Overview of Selected Variables.

Demographics	Age; ICU Length of Stay; Gender; Body Mass Index (BMI); Race; Admission Type (Emergency)
Vital Signs	Heart Rate; Respiratory Rate; Diastolic Blood Pressure (DBP); Systolic Blood Pressure (SBP); Temperature; Oxygen Saturation (SpO_2); Mean Arterial Pressure (MAP)
Laboratory Indicators	Platelet; Prothrombin Time (PT); Partial Thromboplastin Time (PTT); Total Bilirubin; Direct Bilirubin; Lactate Dehydrogenase; Serum Albumin; Blood Urea Nitrogen (BUN); Anion Gap; Base Excess; Alanine Aminotransferase (ALT); Alkaline Phosphatase (ALP); Hematocrit; Serum Chloride; Aspartate Aminotransferase (AST); Creatinine; Blood Glucose; Hemoglobin; White Blood Cell; PaO_2/FiO_2 Ratio; Red Blood Cell; Serum Sodium; Bicarbonate; Serum Potassium; Serum Calcium; Lactate; Neutrophil-to-Lymphocyte Ratio (NLR); C-reactive Protein (CRP); International Normalized Ratio (INR); Magnesium; Fibrinogen
Comorbidities	Hypertension; Congestive Heart Failure; Myocardial Infarction; Liver Disease; Cerebrovascular Disease; Chronic Pulmonary Disease; Renal Disease; Diabetes
Infection Sites	Intestinal Infection; Urinary Infection; Lung Infection; Catheter-Related Infection; Skin and Soft Tissue Infection; Abdominal Cavity Infection; Central Nervous System Infection
Interventions	Vasopressor Use; Mechanical Ventilation; Blood Transfusion
Clinical Measurements	Total Urine Output on Day 1; Glomerular Filtration Rate (GFR)
Severity Scores	GCS Score; SOFA Score; qSOFA Score; APACHE II Score
Outcome	28-day Mortality

2.5. Feature Selection

In this study, the process of feature selection integrated existing literature, clinical relevance, and expert advice. Based on extensive literature and clinical relevance confirmed by domain experts, 68 initial features were extracted from the MIMIC-IV database. These covered demographics, vital signs, laboratory indicators, comorbidities, infection sites, clinical interventions, and severity scores. Each feature was measured within the first 24 h of ICU admission to capture early physiological variation. Both minimum and maximum values were computed for dynamic physiological and biochemical indicators, and 106 features were selected. Prior to final feature selection, data preprocessing was performed to ensure consistency and robustness. In this process, those with more than 20% missing values were excluded, and 88 features remained. The variance inflation factor (VIF) is used to identify and exclude features with high multicollinearity. Features with a VIF value greater than 20 are removed, and 86 features were retained. Subsequently, highly correlated predictors ($|r| > 0.85$) were excluded, resulting in 82 features before model-based selection.

Final feature selection was conducted using permutation feature importance measurement from the best-performing XGBoost model, optimized through grid search with 10-fold cross-validation. The best performing XGBoost model used a column subsampling ratio of 1, a learning rate of 0.1, a maximum tree depth of 5, 200 boosting estimators, and a subsampling ratio of 0.8. Permutation feature importance measurement was computed on the validation set with 10 repeats, and features were ranked according to their mean importance scores. This method estimates feature relevance by quantifying the decrease in model performance after permuting each feature, offering a direct and model-agnostic measure of predictive contribution while mitigating biases present in coefficient-based or tree-based importance measures [22]. The top 30 features were retained for subsequent analyses, including Lactate Min, DiaBP Min, Hemoglobin Max, Total Bilirubin Min, ICU length of stay, White Blood Cell Min, Heart Rate Max, Glucose Serum Max, PTT Max, PTT Min, Platelet Max, Temperature Min, Age, Hematocrit Min, Urine output total, Serum Calcium Min, White Blood Cell Max, APACHE II, BUN Min, Serum Calcium Max, Serum Potassium Min, Respiratory Rate Max, SysBP Min, Temperature Max, Bicarbonate Min, Urinary Infection, Diabetes, Race, Alkaline Phosphatase Max, and GCS score. These features were used to retrain the models, improving both predictive performance and interpretability. The top 30 selected features ranked by mean permutation importance are presented in Figure 2, and their pairwise correlations are illustrated in Figure 3.

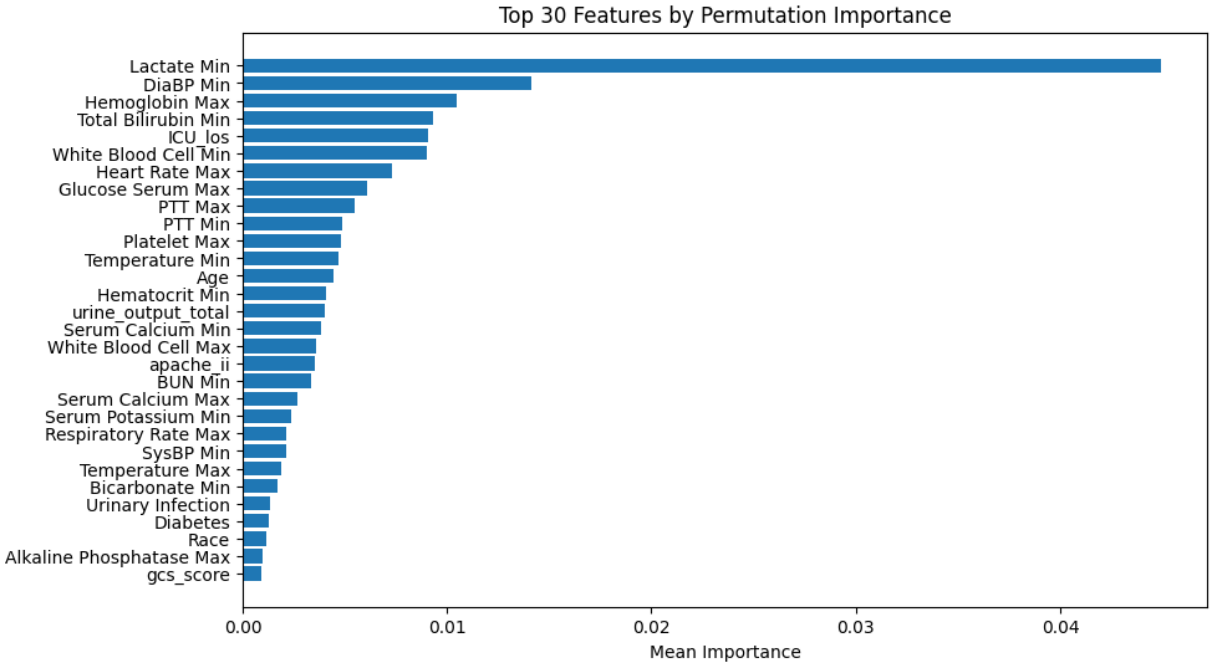


Figure 2. Top 30 Selected Features With Mean Importance.

2.6. Model Development

The dataset was split into training and validation sets in a 7:3 ratio using a train-test split to preserve outcome distribution. Eight machine learning models were evaluated to predict 28-day mortality among patients, including logistic regression, elastic-net penalized logistic regression (cv-glmnet), random forest, support vector machine (SVM), decision tree, k-nearest neighbors (KNN), linear discriminant analysis (LDA), and extreme gradient boosting (XGBoost). All models were developed and validated on the same train–test split, ensuring consistency and fairness in performance comparison. Logistic regression was implemented as the baseline model for its interpretability, while XGBoost was selected as the proposed model due to its capability to handle nonlinear associations and higher-order feature interactions [23,24]. For each model, hyperparameters were optimized using grid

search with stratified 5-fold cross-validation, with the area under the receiver operating characteristic curve (ROC-AUC) serving as the primary scoring metric. The parameter search grids included regularization strength (logistic regression), maximum tree depth, number of estimators, and splitting criteria (tree-based models), kernel functions and penalty parameters (SVM), neighborhood size and distance metrics (kNN), solver choices (LDA), and a comprehensive set of boosting parameters for XGBoost (learning rate, maximum depth, estimators, subsampling ratios, column subsampling ratios, child weight, and gamma). Hyperparameter tuning improved model generalizability and predictive accuracy, ensuring optimal performance for the classification task. Model performance was assessed on the held-out validation set using ROC-AUC with 95% confidence intervals, accuracy, precision, recall, F1-score, and specificity. For each model, the probability threshold was optimized using the precision–recall curve to maximize the F1-score, and confusion matrices were generated to provide a detailed breakdown of classification outcomes.

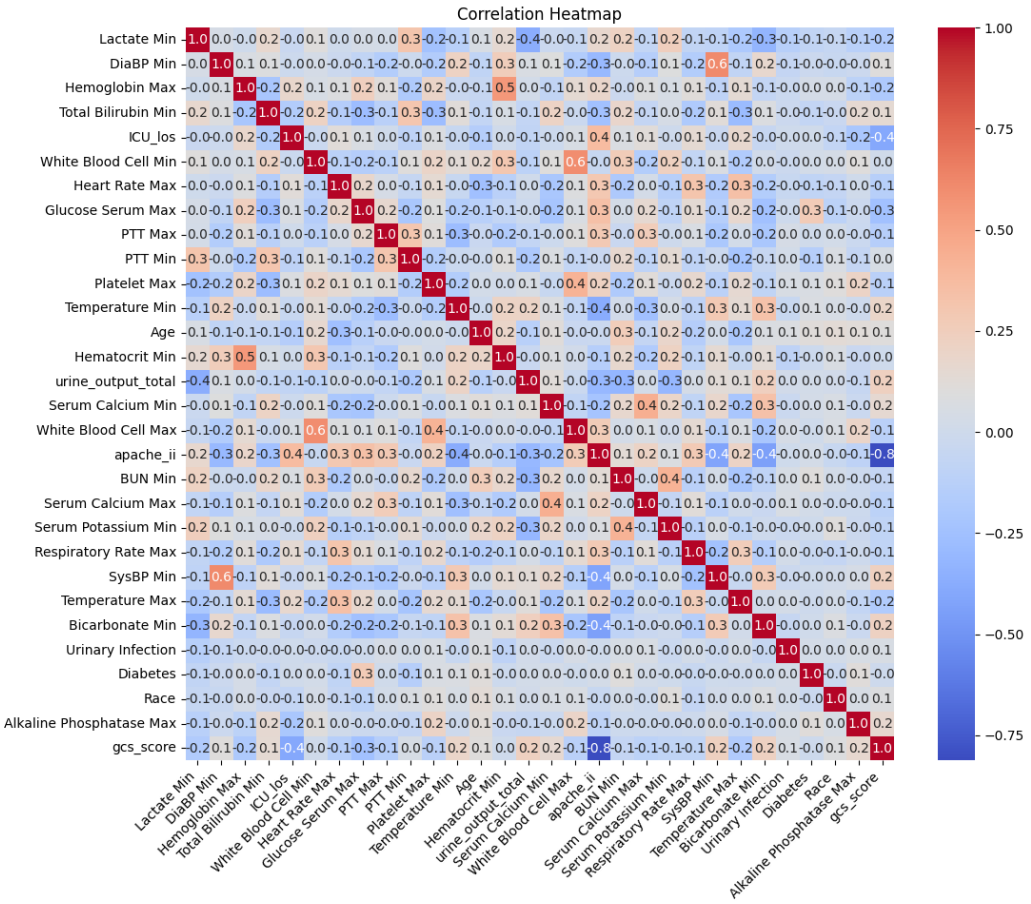


Figure 3. Correlation Heatmap.

2.7. Model Explainability

Model interpretability was addressed using both global and local explanation techniques to improve transparency and clinical applicability. SHapley Additive exPlanations (SHAP), a game-theoretic approach that quantifies the contribution of each feature to individual predictions, was computed for the best-performing model [25,26]. Feature importance was visualized using SHAP summary plots, enabling the identification of variables with the greatest overall influence on 28-day mortality predictions. In addition, Local Interpretable Model-Agnostic Explanations (LIME) were applied to generate instance-level explanations. LIME approximates the complex model with a simple surrogate model within the neighborhood of a single instance, thereby highlighting the relative importance of individual features in that patient's predicted outcome [27]. In this study, LIME was applied to

representative patients from the validation cohort to demonstrate how the top 30 selected predictors influenced individual mortality risk estimates. SHAP and LIME improve the transparency of the machine learning model and facilitate the clinical interpretability of the results.

2.8. Statistical Analysis

To analyze cohort-level variability, two complementary statistical analyses were performed. The first compared survivors and non-survivors to identify features potentially associated with patient outcomes, while the second assessed the comparability of the training and validation datasets to ensure that the training cohort was representative of the validation cohort. Continuous variables were expressed as medians with interquartile ranges (IQR), while categorical variables were assessed using the chi-square test and summarized as percentages. All analyses were conducted in Python (Version 3.10.12). A two-sided p-value < 0.05 was considered statistically significant [28]. This dual approach provides a comprehensive evaluation of both cohort representativeness and clinical outcome differences, establishing a solid foundation for downstream modeling.

3. Results

3.1. Statistical Evaluation

In my study cohort, there were a total of 1157 patients, including 826 patients from the MIMIC-IV database, 225 patients from the MIMIC-III database, and 106 patients from the eICU database. Within the MIMIC-IV database, 578 patients were assigned to the training set and 248 patients to the internal validation set. The records from the MIMIC-III and eICU databases were used as external validation sets. A total of 30 features were ultimately selected for comparison between the training set and the internal validation set from MIMIC-IV, shown in Table 2, as well as between survivors and non-survivors, shown in Table 3. In Table 2, the majority of features had p-values above 0.05, demonstrating no statistically significant differences between the two cohorts. Only Temperature Max showed a modest but statistically significant difference ($p = 0.043$), while all other features were well balanced across cohorts. In contrast, Table 3 indicates that 22 features showed statistically significant differences, with p-values approaching zero or less than 0.05. The exceptions were `ICU_los` ($p = 0.691$), Heart Rate Max ($p = 0.221$), Age ($p = 0.165$), Hematocrit Min ($p = 0.503$), Serum Calcium Min ($p = 0.922$), White Blood Cell Max ($p = 0.801$), Serum Calcium Max ($p = 0.229$), and Alkaline Phosphatase Max ($p = 0.478$). In addition, the comparisons between the training set and the external validation sets are presented in Tables 4 and 5. These comprehensive analyses collectively ensured both the internal consistency of the dataset and the robustness of the models across different patient populations.

Table 2. Baseline Characteristics of Training and Internal Validation Cohorts.

Feature	Training Cohort (N = 578)	Internal Validation Cohort (N = 248)	*p*-Value
Temperature Max	37.50 (37.11, 38.28)	37.71 (37.17, 38.50)	0.043
BUN Min	23.00 (14.00, 38.00)	20.00 (13.00, 32.25)	0.072
Hematocrit Min	25.70 (22.30, 30.10)	24.50 (21.65, 29.82)	0.100
White Blood Cell Min	9.70 (5.80, 14.57)	9.25 (5.05, 14.30)	0.103
Age	66.00 (55.00, 75.00)	63.00 (53.75, 73.00)	0.119
Heart Rate Max	117.00 (101.00, 131.00)	118.00 (102.75, 135.00)	0.279
Respiratory Rate Max	32.00 (28.00, 37.00)	32.00 (28.00, 36.25)	0.312
White Blood Cell Max	18.60 (12.40, 26.50)	17.40 (12.25, 25.35)	0.391
Total Urine Output	1087.50 (552.00, 1860.00)	1087.50 (536.75, 1770.50)	0.402
Serum Potassium Min	3.70 (3.30, 4.10)	3.60 (3.30, 4.10)	0.425

Table 2. *Cont.*

Feature	Training Cohort (N = 578)	Internal Validation Cohort (N = 248)	*p*-Value
Race			0.466
White, n (%)	380 (65.74%)	163 (65.73%)	
Black, n (%)	36 (6.23%)	29 (11.69%)	
Asian, n (%)	13 (2.25%)	9 (3.63%)	
Hispanic, n (%)	22 (3.81%)	7 (2.82%)	
Other, n (%)	127 (21.97%)	40 (16.13%)	
SysBP Min	78.00 (69.00, 87.00)	77.00 (68.00, 85.00)	0.509
Total Bilirubin Min	2.10 (1.00, 4.67)	2.10 (0.80, 4.60)	0.564
Bicarbonate Min	17.00 (14.00, 21.00)	18.00 (14.00, 21.00)	0.581
Alkaline Phosphatase Max	178.50 (105.00, 328.00)	163.50 (105.00, 308.75)	0.590
Hemoglobin Max	10.80 (9.60, 12.20)	10.95 (9.20, 12.43)	0.608
Lactate Min	1.70 (1.20, 2.50)	1.65 (1.20, 2.42)	0.616
PTT Max	45.55 (36.32, 66.97)	45.55 (36.38, 72.45)	0.624
Diabetes			0.647
Yes, n (%)	220 (38.06%)	80 (32.26%)	
No, n (%)	358 (61.94%)	168 (67.74%)	
Temperature Min	36.33 (35.72, 36.50)	36.28 (35.72, 36.50)	0.683
DiaBP Min	39.00 (32.00, 46.00)	41.00 (33.00, 46.00)	0.759
Serum Calcium Max	8.50 (8.00, 9.10)	8.50 (8.00, 9.10)	0.787
Platelet Max	183.00 (115.00, 281.00)	176.00 (103.75, 283.50)	0.825
GCS Score	10.00 (3.00, 15.00)	9.00 (3.00, 14.00)	0.837
ICU Length of Stay	92.58 (48.21, 232.23)	95.89 (51.29, 215.29)	0.853
PTT Min	32.45 (28.00, 38.68)	32.48 (28.88, 38.12)	0.865
Serum Calcium Min	7.50 (7.00, 8.00)	7.50 (7.00, 8.00)	0.919
Serum Glucose Max	163.00 (125.00, 240.00)	160.00 (126.75, 226.25)	0.930
Apache II	23.00 (16.00, 30.00)	22.00 (16.00, 30.00)	0.932
Urinary Infection			0.960
Yes, n (%)	176 (30.45%)	77 (31.05%)	
No, n (%)	402 (69.55%)	171 (68.95%)	

Table 3. Baseline Characteristics of Survivors and Non-survivors.

Feature	Total (N = 826)	Survivors (N = 500)	Non-Survivors (N = 326)	*p*-Value
Lactate Min	1.60 [1.20–2.60]	1.40 [1.10–2.10]	2.20 [1.50–3.80]	0.000
DiaBP Min	40 [33–46]	40.00 [34.00–46.00]	38.00 [30.00–46.00]	0.016
Hemoglobin Max	10.90 [9.40–12.30]	11.20 [9.80–12.60]	10.30 [9.00–11.80]	0.000
Total Bilirubin Min	2.10 [0.90–4.90]	1.70 [0.70–3.80]	2.90 [1.30–8.33]	0.000
ICU Length of Stay	93.15 [48.92–222.59]	92.99 [48.76–250.22]	93.47 [49.21–204.75]	0.691
White Blood Cell Min	9.60 [5.60–14.50]	8.90 [5.50–13.10]	10.80 [5.90–16.40]	0.002
Heart Rate Max	117 [101–132]	116.00 [100.00–131.25]	118.00 [102.00–133.00]	0.221
Serum Glucose Max	163 [126–234]	167.00 [131.00–239.50]	156.00 [120.00–223.00]	0.033
PTT Max	45.60 [36.10–69.00]	43.70 [34.90–64.30]	50.50 [39.08–78.10]	0.000
PTT Min	32.50 [28.10–38.70]	31.00 [27.30–35.80]	35.50 [30.50–44.30]	0.000
Platelet Max	182 [112–282]	197.50 [127.75–291.50]	153.00 [87.00–251.00]	0.000
Temperature Min	36.33 [35.72–36.51]	36.33 [35.89–36.56]	36.22 [35.56–36.50]	0.000
Age	65 [55–75]	65.00 [53.00–76.00]	66.00 [57.00–75.00]	0.165
Hematocrit Min	25.30 [22.00–30.00]	25.55 [22.00–30.12]	25.00 [22.00–29.40]	0.503
Total Urine Output	1090 [524.25–1863.75]	1286 [770.25–2225]	750 [271.75–1400]	0.000
Serum Calcium Min	7.50 [7.00–8.00]	7.50 [7.00–8.00]	7.50 [6.90–8.10]	0.922
White Blood Cell Max	18.20 [12.30–26.40]	18.10 [12.40–26.02]	18.90 [11.80–26.50]	0.801

Table 3. *Cont.*

Feature	Total (N = 826)	Survivors (N = 500)	Non-Survivors (N = 326)	*p*-Value
Apache II	22.00 [16.00–30.00]	21.00 [15.00–29.00]	25.00 [18.00–31.00]	0.000
BUN Min	22.00 [14.00–37.00]	19.00 [12.00–31.00]	29.00 [18.00–46.00]	0.000
Serum Calcium Max	8.50 [8.00–9.10]	8.50 [8.00–9.10]	8.40 [7.90–9.10]	0.229
Serum Potassium Min	3.70 [3.30–4.10]	3.60 [3.30–4.00]	3.80 [3.40–4.40]	0.000
Respiratory Rate Max	32.00 [28.00–37.00]	32.00 [28.00–37.00]	31.00 [27.00–36.00]	0.017
SysBP Min	78.00 [68.00–86.00]	78.00 [70.00–87.00]	77.00 [66.00–84.00]	0.021
Temperature Max	37.58 [37.11–38.33]	37.72 [37.22–38.50]	37.30 [36.94–38.06]	0.000
Bicarbonate Min	18.00 [14.00–21.00]	18.00 [15.00–21.00]	17.00 [13.00–21.00]	0.011
Alkaline Phosphatase Max	172.00 [103.00–329.50]	177.00 [105.00–329.00]	166.00 [100.25–329.50]	0.478
GCS Score	10.00 [3.00–14.00]	10.00 [3.00–15.00]	8.00 [3.00–14.00]	0.001
Urinary Infection				0.007
Yes, n (%)	253 (30.6%)	171 (34.2%)	82 (25.2%)	
No, n (%)	573 (69.4%)	329 (65.8%)	244 (74.8%)	
Diabetes				0.000
Yes, n (%)	300 (36.3%)	210 (42.0%)	90 (27.6%)	
No, n (%)	526 (63.7%)	290 (58.0%)	236 (72.4%)	
Race				0.000
White, n (%)	543 (65.7%)	344 (68.8%)	199 (61.0%)	
Black, n (%)	65 (7.9%)	45 (9.0%)	20 (6.1%)	
Asian, n (%)	22 (2.7%)	17 (3.4%)	5 (1.5%)	
Hispanic, n (%)	29 (3.5%)	18 (3.6%)	11 (3.4%)	
Other, n (%)	167 (20.2%)	76 (15.2%)	91 (27.9%)	

Table 4. Comparison of Training and External Validation Cohort (MIMIC-III).

Feature	Train Cohort (N = 578)	External Validation Cohort (N = 225)	*p*-Value
GCS Score	10.00 (3.00, 15.00)	0.00 (0.00, 7.00)	0
Apache II	23.00 (16.00, 30.00)	27.00 (22.00, 33.00)	<0.001
Temperature Min	36.33 (35.72, 36.50)	35.67 (35.17, 36.28)	<0.001
Hemoglobin Max	10.80 (9.60, 12.20)	11.40 (10.10, 12.70)	<0.001
Total Bilirubin Min	2.10 (1.00, 4.67)	3.20 (1.50, 7.10)	0.002
SysBP Min	78.00 (69.00, 87.00)	78.00 (66.00, 86.00)	0.002
Age	66.00 (55.00, 75.00)	63.00 (51.00, 77.00)	0.006
Lactate Min	1.70 (1.20, 2.50)	1.70 (1.30, 2.90)	0.011
PTT Min	32.45 (28.00, 38.68)	34.20 (29.80, 41.90)	0.013
Hematocrit Min	25.70 (22.30, 30.10)	26.90 (23.80, 31.60)	0.022
ICU Length of Stay	92.58 (48.21, 232.23)	103.60 (54.07, 263.99)	0.031
BUN Min	23.00 (14.00, 38.00)	26.00 (15.00, 40.00)	0.092
Temperature Max	37.50 (37.11, 38.28)	37.72 (37.17, 38.44)	0.099
Serum Calcium Min	7.50 (7.00, 8.00)	7.40 (6.70, 7.80)	0.103
DiaBP Min	39.00 (32.00, 46.00)	39.00 (32.00, 45.00)	0.115
Bicarbonate Min	17.00 (14.00, 21.00)	17.00 (13.00, 20.00)	0.136
Alkaline Phosphatase Max	178.50 (105.00, 328.00)	185.00 (115.00, 310.00)	0.249
Respiratory Rate Max	32.00 (28.00, 37.00)	31.00 (27.00, 36.00)	0.281
Heart Rate Max	117.00 (101.00, 131.00)	117.00 (104.00, 133.00)	0.358
PTT Max	45.55 (36.32, 66.97)	47.90 (38.60, 70.30)	0.501
Platelet Max	183.00 (115.00, 281.00)	185.00 (106.00, 301.00)	0.546
Serum Potassium Min	3.70 (3.30, 4.10)	3.60 (3.30, 4.00)	0.590
Urinary Infection			0.607
Yes, n (%)	176 (30.45%)	49 (21.78%)	
No, n (%)	402 (69.55%)	176 (78.22%)	
Serum Calcium Max	8.50 (8.00, 9.10)	8.40 (7.90, 9.10)	0.618

Table 4. *Cont.*

Feature	Train Cohort (N = 578)	External Validation Cohort (N = 225)	*p*-Value
Diabetes			0.645
Yes, n (%)	220 (38.06%)	69 (30.67%)	
No, n (%)	358 (61.94%)	156 (69.33%)	
White Blood Cell Min	9.70 (5.80, 14.57)	9.80 (6.40, 15.40)	0.665
Race			0.931
White, n (%)	380 (65.74%)	150 (66.67%)	
Black, n (%)	36 (6.23%)	17 (7.56%)	
Asian, n (%)	13 (2.25%)	5 (2.22%)	
Hispanic, n (%)	22 (3.81%)	12 (5.33%)	
Other, n (%)	127 (21.97%)	41 (18.22%)	
Serum Glucose Max	163.00 (125.00, 240.00)	163.00 (119.00, 227.00)	0.940
Total Urine Output	1087.50 (552.00, 1860.00)	985.00 (363.00, 1730.00)	0.948
White Blood Cell Max	18.60 (12.40, 26.50)	18.10 (12.40, 26.60)	0.990

Table 5. Comparison of Training and External Validation Cohorts (eICU-CRD).

Feature	Train Cohort (N = 578)	External Validation Cohort (N = 106)	*p*-Value
Lactate Min	1.70 (1.20, 2.50)	2.50 (1.70, 4.97)	<0.001
PTT Min	32.45 (28.00, 38.68)	35.50 (32.45, 48.90)	<0.001
Platelet Max	183.00 (115.00, 281.00)	134.00 (79.25, 192.75)	<0.001
Age	66.00 (55.00, 75.00)	59.00 (50.25, 66.75)	<0.001
PTT Max	45.55 (36.32, 66.97)	45.55 (41.00, 52.00)	0.002
Serum Calcium Max	8.50 (8.00, 9.10)	8.30 (7.60, 8.80)	0.005
Serum Glucose Max	163.00 (125.00, 240.00)	141.50 (115.25, 193.75)	0.009
Total Bilirubin Min	2.10 (1.00, 4.67)	3.15 (1.30, 7.38)	0.011
BUN Min	23.00 (14.00, 38.00)	32.00 (17.25, 47.75)	0.017
White Blood Cell Max	18.60 (12.40, 26.50)	15.52 (10.83, 21.80)	0.018
Respiratory Rate Max	32.00 (28.00, 37.00)	33.00 (30.00, 43.50)	0.024
SysBP Min	78.00 (69.00, 87.00)	78.00 (74.25, 78.00)	0.077
Temperature Max	37.50 (37.11, 38.28)	37.50 (37.50, 37.50)	0.107
GCS Score	10.00 (3.00, 15.00)	11.00 (6.25, 15.00)	0.150
DiaBP Min	39.00 (32.00, 46.00)	39.00 (39.00, 39.00)	0.152
Serum Potassium Min	3.70 (3.30, 4.10)	3.70 (3.30, 4.10)	0.174
ICU Length of Stay	92.58 (48.21, 232.23)	121.58 (59.75, 241.87)	0.245
Heart Rate Max	117.00 (101.00, 131.00)	117.50 (104.00, 133.00)	0.416
Serum Calcium Min	7.50 (7.00, 8.00)	7.45 (6.80, 8.10)	0.484
Hematocrit Min	25.70 (22.30, 30.10)	26.90 (22.02, 30.80)	0.542
Apache II	23.00 (16.00, 30.00)	23.00 (18.00, 28.75)	0.609
Alkaline Phosphatase Max	178.50 (105.00, 328.00)	161.00 (103.00, 295.50)	0.629
Temperature Min	36.33 (35.72, 36.50)	36.33 (36.33, 36.33)	0.652
Total Urine Output	1087.50 (552.00, 1860.00)	1047.50 (375.25, 1511.00)	0.720
Diabetes			0.744
Yes, n (%)	220 (38.06%)	9 (8.49%)	
No, n (%)	358 (61.94%)	97 (91.51%)	
Urinary Infection			0.804
Yes, n (%)	176 (30.45%)	15 (14.15%)	
No, n (%)	402 (69.55%)	91 (85.85%)	
Hemoglobin Max	10.80 (9.60, 12.20)	10.50 (8.93, 12.67)	0.874
Bicarbonate Min	17.00 (14.00, 21.00)	17.00 (14.00, 20.00)	0.893
White Blood Cell Min	9.70 (5.80, 14.57)	10.60 (6.40, 15.15)	0.937
Race			0.998
White, n (%)	380 (65.74%)	78 (73.58%)	
Black, n (%)	36 (6.23%)	8 (7.55%)	
Asian, n (%)	13 (2.25%)	4 (3.77%)	
Hispanic, n (%)	22 (3.81%)	2 (1.89%)	
Other, n (%)	127 (21.97%)	14 (13.21%)	

3.2. Model Performance

Model performance was evaluated across 8 machine learning models for 28-day mortality, with results presented in Table 6. Among the models, XGBoost achieved the highest overall performance with an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.8556 (95% CI: 0.807–0.898), an accuracy of 0.7702, a recall of 0.8469, and a specificity of 0.7200. The optimal hyperparameter configuration for XGBoost included a maximum tree depth of 7, a minimum child weight of 1, 200 estimators, and a learning rate of 0.1, yielding strong performance in terms of accuracy, recall, and generalization. A column subsample ratio of 1.0 and a row subsample ratio of 0.8 were applied to reduce variance, while a gamma value of 0.01 regularized tree splitting. Random Forest and SVM achieved comparable performance with AUROCs of 0.8280 and 0.8098, respectively. The ROC curves for all the models are illustrated in Figure 4. In the external validation analysis, the models demonstrated varying performance across datasets. For the MIMIC-III cohort, the Random Forest model achieved the best performance, with an AUROC of 0.8964 (95% CI: 0.845–0.943), surpassing the XGBoost model identified as optimal in the internal validation. In contrast, for the eICU cohort, XGBoost remained the top-performing model, but with a reduced AUROC of 0.6966 (95% CI: 0.594–0.790). The comparison for external validations is shown in Table 7. The better performance in the MIMIC-III dataset is due to its closer similarity to MIMIC-IV in patient characteristics and data collection, while the multicenter nature of eICU introduces more variability, reducing model generalizability. These findings underscore the importance of external validation in evaluating model robustness and highlight limitations in the model's generalizability across diverse patient populations and clinical settings. The ROC curves of external validation for MIMIC-III and eICU-CRD are presented in Figures 5 and 6, respectively.

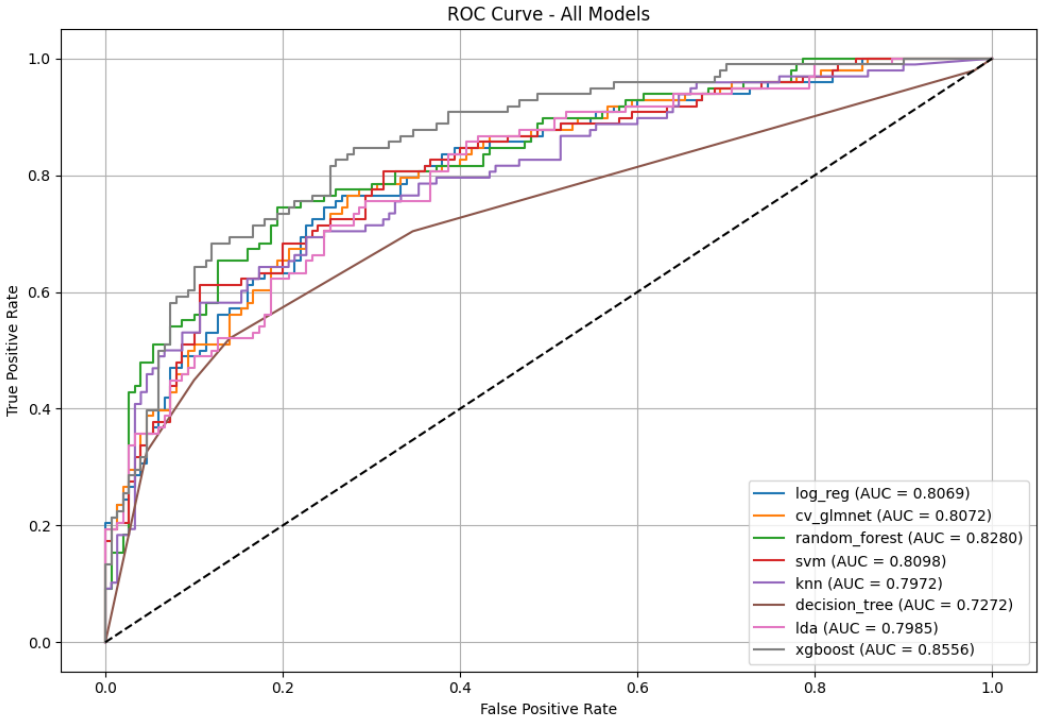


Figure 4. ROC Curves For Internal Validation.

Table 6. Model Performance For Internal Validation.

Model	AUROC	95% CI	Accuracy	Precision	Recall	F1-Score	Specificity
XGBoost	**0.8556**	[0.807, 0.898]	0.7702	0.6640	0.8469	0.7444	0.7200
Random Forest	0.8280	[0.777, 0.879]	0.7823	0.7157	0.7449	0.7300	0.8067
SVM	0.8098	[0.751, 0.868]	0.7339	0.6270	0.8061	0.7054	0.6867
CV_Glmnet	0.8072	[0.753, 0.864]	0.7419	0.6466	0.7653	0.7009	0.7267
Logistic Regression	0.8069	[0.751, 0.864]	0.7460	0.6522	0.7653	0.7042	0.7333
LDA	0.7985	[0.741, 0.855]	0.6975	0.5793	0.8571	0.6914	0.5933
KNN	0.7972	[0.740, 0.853]	0.7419	0.6667	0.6939	0.6800	0.7733
Decision Tree	0.7272	[0.663, 0.785]	0.6734	0.5702	0.7041	0.6301	0.6533

Table 7. Model Performance for External Validations.

Dataset	Model	AUROC	95% CI	Accuracy	Precision	Recall	F1-Score	Specificity
MIMIC-III	**Random Forest**	**0.8964**	**[0.845, 0.943]**	0.8711	0.8219	0.7895	0.8054	0.9128
	XGBoost	0.8800	[0.822, 0.932]	0.8622	0.8082	0.7763	0.7919	0.9060
	KNN	0.8592	[0.805, 0.909]	0.8089	0.6897	0.7895	0.7362	0.8188
	SVM	0.8307	[0.777, 0.885]	0.7556	0.6105	0.7632	0.6784	0.7517
	CV_Glmnet	0.8167	[0.755, 0.875]	0.7511	0.5943	0.8289	0.6923	0.7114
	Logistic Regression	0.8131	[0.752, 0.872]	0.7422	0.5882	0.7895	0.6742	0.7181
	Decision Tree	0.8045	[0.743, 0.856]	0.6800	0.5149	0.9079	0.6571	0.5638
	LDA	0.8026	[0.737, 0.863]	0.7200	0.5556	0.8553	0.6736	0.6510
eICU-CRD	**XGBoost**	**0.6966**	**[0.594, 0.790]**	0.6038	0.5730	0.9273	0.7083	0.2549
	SVM	0.6877	[0.585, 0.779]	0.6038	0.5714	0.9455	0.7123	0.2353
	CV_Glmnet	0.6827	[0.578, 0.780]	0.5660	0.5484	0.9273	0.6892	0.1765
	Logistic Regression	0.6813	[0.580, 0.774]	0.5755	0.5532	0.9455	0.6980	0.1765
	Random Forest	0.6781	[0.567, 0.777]	0.6321	0.6000	0.8727	0.7111	0.3725
	LDA	0.6645	[0.559, 0.763]	0.5377	0.5288	1.0000	0.6918	0.0392
	Decision Tree	0.6414	[0.536, 0.736]	0.5377	0.5300	0.9636	0.6839	0.0784
	KNN	0.6381	[0.540, 0.741]	0.6038	0.5765	0.8909	0.7000	0.2941

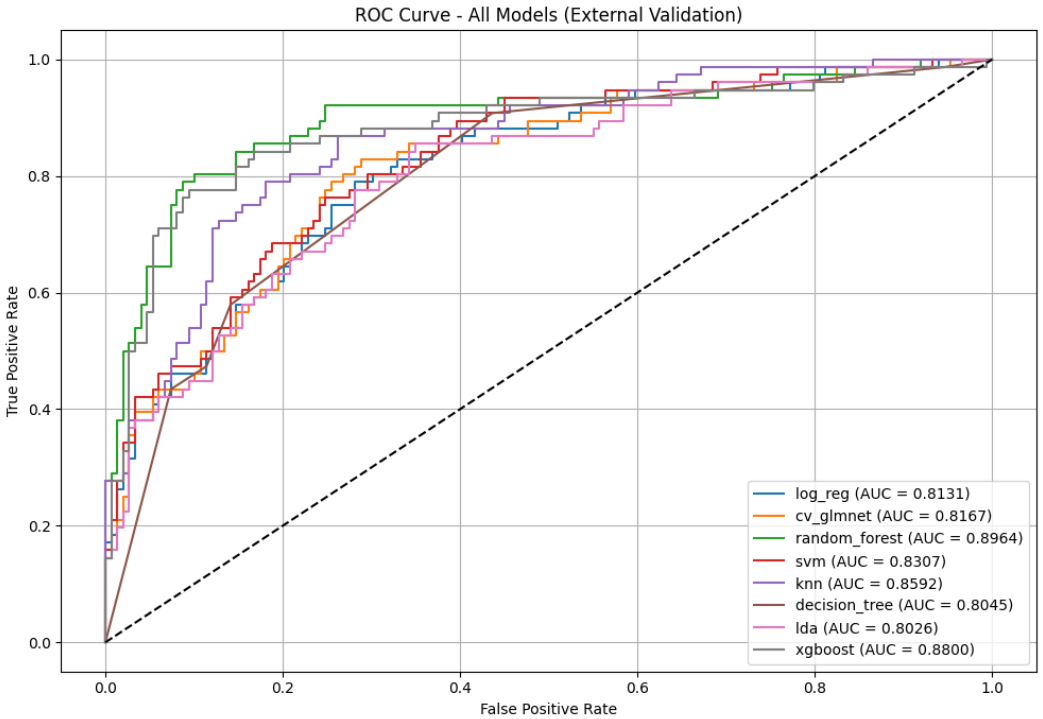


Figure 5. ROC Curves For External Validation (MIMIC-III).

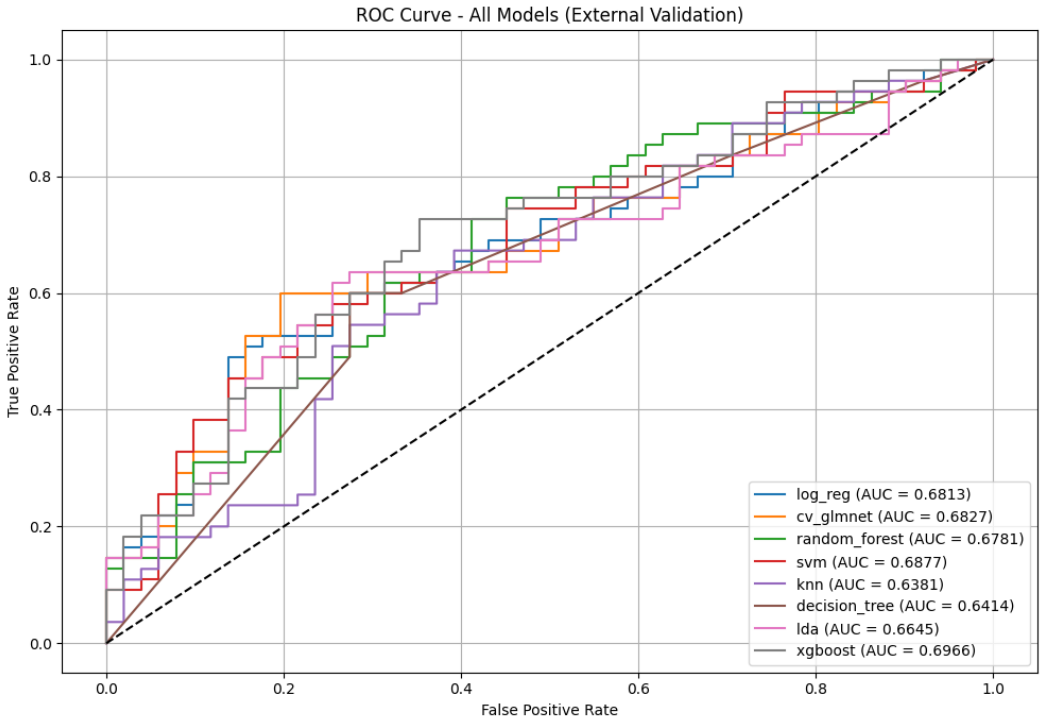


Figure 6. ROC Curves For External Validation (eICU-CRD).

3.3. SHapley Additive exPlanations (SHAP) Value Analysis

In machine learning, each feature is assigned an importance value representing its contribution to the model's output. SHapley Additive exPlanations (SHAP) analysis was applied to the test dataset to evaluate the relative contribution of individual features to the 28-day mortality prediction model in SALI patients. The results highlighted Lactate Min, BUN Min, Heart Rate Max, Hemoglobin Max, and DiaBP Min as the most influential predictors, which were strongly associated with an increased mortality risk. Other important features included DiaBP Min, PTT Max, and Temperature Min, all of which also had a substantial impact on model predictions. Figure 7 presents the SHAP summary plot for the best-performing XGBoost model, where red dots indicate higher feature values and blue dots indicate lower values, illustrating how variations in feature values are related to either increased or reduced mortality risk. For example, elevated lactate levels, increased blood urea nitrogen levels, rapid heart rate, and decreased hemoglobin levels were key factors contributing to poor prognosis, further confirming their role as indicators of disease severity. Similarly, patients with relatively higher hemoglobin levels or body temperature maintained within the normal range demonstrated lower predicted mortality risk. These findings provide clinicians with evidence-based insights into protective factors, which can inform supportive treatment strategies and risk management measures. SHAP analysis enhances transparency in model predictions, allowing clinicians to better understand the sources of predicted risk for each patient. This transparency not only increases trust in the model but also supports clinical decision-making, enabling clinicians to integrate the model's predictions with their own experience and judgment when developing treatment plans.

3.4. Local Interpretable Model-Agnostic Explanations (LIME) Analysis

Local Interpretable Model-Agnostic Explanations (LIME) analysis was conducted to provide local interpretability of the best-performing XGBoost model by explaining the prediction for an individual patient classified as high risk of 28-day mortality. As shown in Figure 8, several features demonstrated strong contributions to the predicted outcome. Green bars represent features that increase the risk of death, whereas red bars indicate

features that reduce the risk. The top contributors to increased mortality risk included Lactate Min (>0.44), BUN Min (>0.69), Serum Potassium Min (>0.53), Total Bilirubin Min (>0.54), and SysBP Min (>0.59). In contrast, protective factors such as Heart Rate Max (≤ -0.75), Temperature Max (≤ -0.74), `urine_output_total` (≤ -0.34), PTT Max (-0.75 to -0.29), `ICU_los` (≤ -0.78), and DiaBP Min (≤ 0.65) lowered the predicted risk. Notably, these threshold values are based on standardized feature values after preprocessing and normalization, so negative values indicate relative deviations from the cohort mean rather than actual physiological measurements. This analysis provided case-level interpretability, demonstrating how specific feature values influenced the predicted mortality risk for an individual patient and highlighting its potential in clinical settings. For example, patients with markedly elevated lactate and BUN levels may benefit from prioritized circulatory support and renal protection, while concurrent reductions in urine output and body temperature may highlight the need for closer monitoring and early intervention. These contributions enhance the clinical applicability of the model by providing actionable, interpretable, and patient-centered insights. The LIME analysis enables clinicians to perform individualized risk stratification based on each patient's clinical profile.

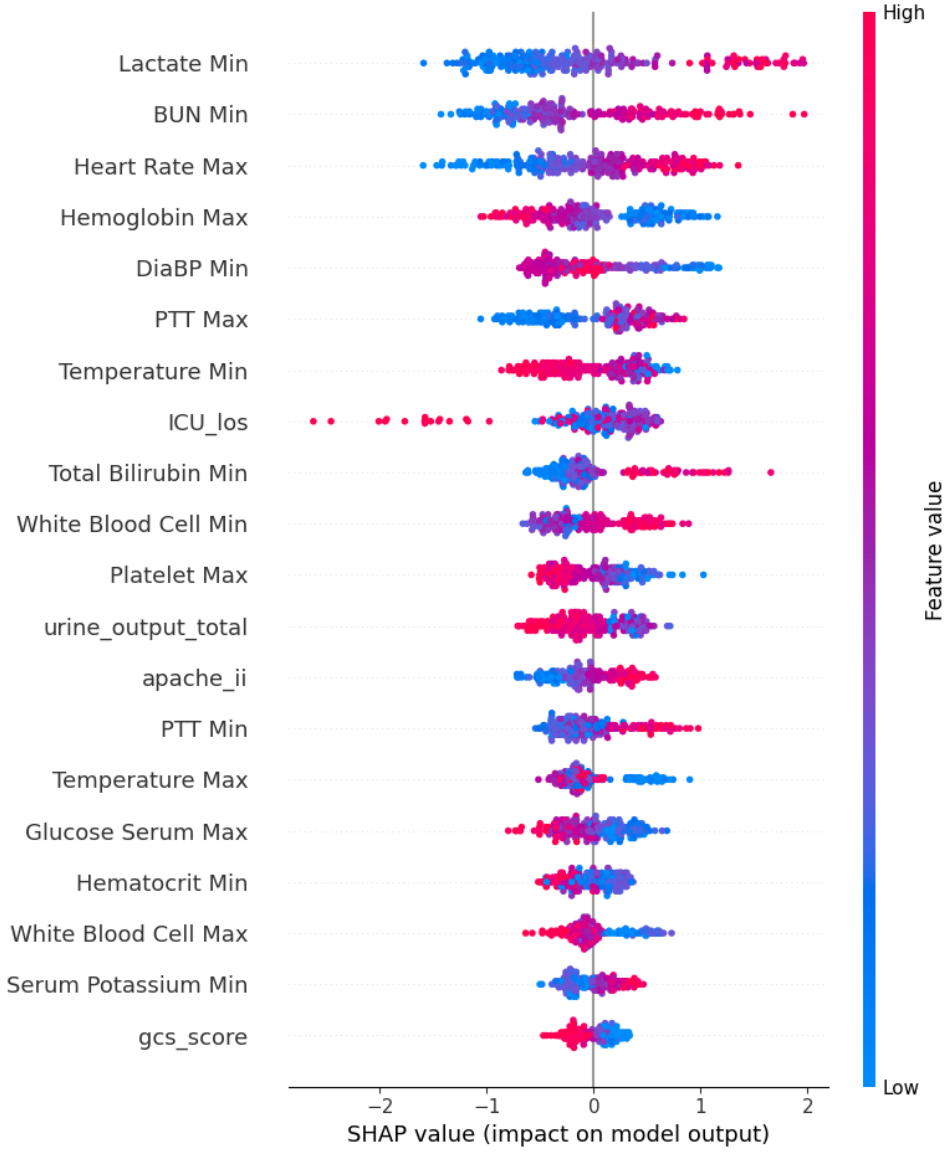


Figure 7. SHAP Summary Plot.

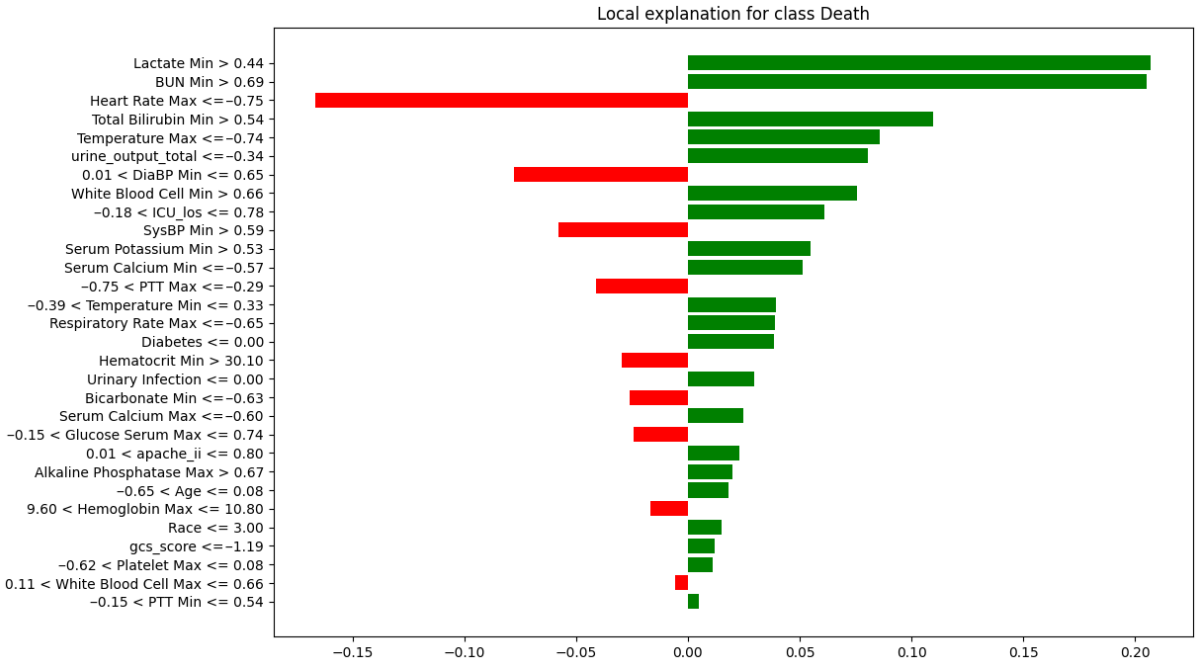


Figure 8. LIME Local Explanation Plot.

4. Discussion

4.1. Comparison with Existing Literature

Wen et al. [15] developed a Random Forest model for 28-day mortality prediction, which reported an AUROC of 0.79 (95% CI: 0.73–0.86), making it the best-performing model in the existing literature for this task. Their model achieved an accuracy of 0.7465, precision of 0.7552, recall of 0.8500, and specificity of 0.5930 [15]. However, our XGBoost model demonstrated superior predictive capability, achieving an AUROC of 0.8556 (95% CI: 0.807–0.898). This represents an approximate 8.3% improvement compared with the AUROC reported by Wen et al. (2024) [15], underscoring its stronger discriminatory power. In addition, the XGBoost model achieved higher accuracy (0.7702) and specificity (0.7200) than the Random Forest model's values of 0.7465 and 0.5930, respectively. These results indicate that this method not only improves the reliability of identifying high-risk patients but also reduces the likelihood of misclassifying low-risk individuals, thereby minimizing unnecessary alerts and resource use. Although Wen et al.'s model achieved slightly higher precision of 0.7552 compared with 0.6640 and recall of 0.8500 compared with 0.8469, these gains were offset by a weaker balance between sensitivity and specificity. By contrast, the XGBoost model demonstrated more consistent performance across evaluation metrics, particularly AUROC and specificity, reinforcing its clinical utility in mortality prediction. Our objective was to prioritize a clinically meaningful balance between sensitivity and specificity over the maximization of any single performance metric, thereby ensuring the reliable detection of high-risk cases while mitigating the risk of alert fatigue. Importantly, the model demonstrated enhanced sensitivity that reflects a superior ability to detect true high-risk cases, which remains a critical factor for successful early interventions in the intensive care unit. By achieving a superior balance between sensitivity and specificity through XGBoost, our study highlights the value of modeling non-linear interactions among physiological markers. These interactions provide a more stable and clinically informative predictive signal for sepsis-associated liver injury compared with approaches that rely more heavily on demographic or coarse clinical predictors used in prior random forest models. A comparison between the proposed XGBoost model and Wen et al.'s model

is presented in Table 8. For interpretability, SHAP analysis was applied to identify key predictive features in a prior study [15].

Table 8. Comparison of Performance Metrics.

Work	Best Model	AUROC [95% CI]	Accuracy	Precision	Recall	Specificity
Our study	XGBoost	**0.8556 [0.807–0.898]**	0.7702	0.6640	**0.8469**	**0.7200**
Wen et al. (2024) [15]	Random Forest	0.7900 [0.73–0.86]	**0.7465**	**0.7552**	0.8500	0.5930

Instead of only using the SHAP analysis, our study also integrated LIME analysis, which allows for both global and individual-level interpretation of results, thus enhancing its clinical application value. This dual-layer interpretability framework addresses a primary challenge in clinical artificial intelligence by facilitating the transition from global-level statistical insights to individual-level decision support tailored to the unique physiological profiles of specific patients. SHAP and LIME analysis identified lactate, blood urea nitrogen, hemoglobin, and heart rate as the most influential predictors, which align well with the established clinical understanding of sepsis-associated liver injury and its impact on patient outcomes [29,30]. Our model identified risk factors that align with what clinicians already recognize as important in sepsis care, which enhances the credibility and clinical trustworthiness of the model.

In addition, Wen et al.'s (2024) [15] study was limited to using a single database, MIMIC-III, as external validation, which limited the generalizability of their findings. Our study incorporated multi-center external validation using both the MIMIC-III and eICU databases, thereby providing stronger evidence for the model's robustness and generalizability. By identifying the generalization gap across these distinct cohorts, our study emphasized how institutional heterogeneity necessitates formal domain adaptation when applying models trained on single-center datasets to broader clinical environments. While class imbalance was less severe in our cohort, we applied stratified sampling to preserve outcome distributions and minimize bias during model training.

Our approach addresses several key limitations in prior studies, particularly in feature selection and preprocessing. In Wen et al.'s (2024) [15] study, feature selection was limited to screening a fixed set of 24 variables, and model construction relied on direct coefficient interpretation, where positive coefficients indicated higher mortality risk and negative coefficients indicated protective effects. While this approach provides straightforward interpretability, it may overlook non-linear interactions and complex dependencies between features. In contrast, our study employed a more comprehensive feature engineering and selection framework that combined statistical testing, clinical expertise, and model-based selection methods (e.g., SHAP values and permutation importance) to identify the top 30 predictors of 28-day mortality. This process reduced the risk of bias and ensured that only clinically meaningful and predictive variables were included. The inclusion of 30 features, compared to 24 in Wen et al.'s study [15], enabled the model to capture additional non-linear relationships and subtle physiological variations that might be overlooked in simpler models. Notably, the top five key predictors in our model (Lactate Min, BUN Min, Heart Rate Max, Hemoglobin Max, and DiaBP Min) differ substantially from those identified by Wen et al., where urine output, Charlson Comorbidity Index, GCS score Min, BUN, and admission age dominated the feature importance ranking. Except for BUN, our predictors focus more directly on hemodynamic and metabolic dysfunction, which are core pathophysiological features of sepsis-associated liver injury.

To improve data quality, we removed highly collinear features and standardized continuous variables, thereby strengthening model robustness. These differences highlight the ability of our model to capture clinically meaningful predictors specific to sepsis-associated

liver injury, thereby improving its relevance and applicability for patient outcome prediction. These methodological improvements enabled the model to better capture the disease-specific physiological dynamics unique to SALI. Importantly, the integration of SHAP and LIME enhances transparency, bridging the gap between data-driven predictions and clinical interpretability, thereby supporting evidence-based decision-making in critical care.

4.2. Interpretable Machine Learning in Liver Injury

Interpretable machine learning has been increasingly applied to liver injury research, reflecting growing emphasis on transparency, clinical trust, and ethical deployment of artificial intelligence in the medical field. In recent years, there has been a significant increase in research dedicated to improving the interpretability of machine learning models used in medical diagnosis and risk stratification. Previous studies have shown that interpretable models can support clinical risk stratification by clarifying how individual variables influence model predictions, thereby improving model interpretability and clinician acceptance of these models. For example, prior studies have successfully applied interpretable frameworks to predict Drug-Induced Liver Injury (DILI), identifying specific toxicological triggers in stable patient populations [31,32].

Although these studies have laid the foundation for transparent research into liver injury, there are still significant shortcomings in applying these methods to the high-risk, dynamic environment of sepsis-associated liver injury (SALI). Sepsis-associated liver injury (SALI) differs fundamentally from drug-induced liver injury (DILI) in terms of underlying pathophysiological mechanisms, clinical trajectory, and risk determinants. DILI is usually caused by the direct hepatotoxic effects of medications and typically occurs in patient populations with relatively stable conditions and more predictable drug exposure patterns [31,32]. In contrast, SALI occurs against a backdrop of systemic inflammation, hemodynamic instability, immune dysfunction, and multi-organ dysfunction, all of which contribute to rapid disease progression and complex, variable clinical manifestations [15]. These differences necessitate more robust modeling strategies and stronger interpretability requirements to support clinical decision-making in intensive care units.

In addition, although the studies by Xiao et al. (2024) [31] and Zhong et al. (2021) [32] demonstrated the feasibility of interpretable machine learning in predicting liver injury, their research focused on single-center cohorts and lacked external validation datasets to evaluate model generalizability. In contrast, our study explicitly evaluated the model's performance across multiple intensive care databases, thus providing a more rigorous assessment of the model's generalizability across different clinical settings. Furthermore, previous studies on liver injury by Xiao et al. (2024) [31] and Zhong et al. (2021) [32] typically relied on a single interpretability method or a limited set of features, which may be insufficient to capture the complex non-linear relationships in critically ill patient populations. Our study, by integrating SHAP and LIME methods, provides complementary global and patient-specific explanations, thereby enhancing transparency and supporting individualized risk assessment. This dual-layer interpretability framework aligns with emerging ethical and clinical expectations for applying interpretable machine learning in critical applications. Our research bridges the gap between advanced data science and clinical critical care, providing a transparent and evidence-based tool for mortality risk stratification in intensive care units.

4.3. Interpretability

In our analysis, we identified the five most influential clinical indicators that impact the model's predictive results, including Lactate Min, BUN Min, Heart Rate Max, Hemoglobin

Max, and DiaBP Min, which can serve as the basis for early warning of mortality risk in SALI patients.

In our machine learning model, Lactate Min showed the highest predictive contribution. Elevated lactate reflects impaired tissue perfusion and oxygen metabolism, and is strongly linked to disease severity and mortality in sepsis [33]. Analyses from the Surviving Sepsis Campaign demonstrated that lactate $\geq$ 4 mmol/L is independently associated with increased mortality [34]. As the liver is central to metabolism, patients with liver disease, particularly chronic liver conditions, often present with elevated lactate [35], and arterial levels > 2 mmol/L correlate with organ dysfunction and mortality [36]. In SHAP and LIME analyses, Lactate Min consistently emerged as a dominant predictor, with values > 0.44 contributing significantly to mortality risk, aligning with clinical evidence that elevated lactate reflects tissue hypoperfusion and impaired hepatic clearance, both hallmarks of sepsis-associated liver injury.

For patients with sepsis-associated liver injury (SALI), blood urea nitrogen (BUN) serves as a key marker reflecting both renal dysfunction and impaired hepatic nitrogen metabolism [37]. Elevated BUN is strongly associated with poor outcomes in sepsis, as it indicates systemic metabolic derangement and multi-organ failure risk [38]. The liver's central role in the urea cycle means that hepatic dysfunction in SALI can reduce ammonia clearance, while concurrent renal impairment further drives BUN elevation [39]. In our SHAP and LIME analyses, BUN Min > 0.69 was a major contributor to mortality risk, supporting its utility as an early warning biomarker.

Moreover, hemoglobin plays a vital role as an oxygen carrier and a crucial factor affecting tissue perfusion. Low hemoglobin levels are closely associated with poor prognosis, as anemia exacerbates tissue hypoxia and liver damage. In sepsis patients, anemia is often caused by inflammation, bone marrow suppression, and hemodilution, and has been identified as an independent predictor of mortality [40,41]. One study showed that in septic shock patients, hemoglobin levels below 9.0 g/dL were associated with higher mortality, and a greater decline in hemoglobin levels was associated with a worse prognosis [42]. In our SHAP and LIME analyses, a decrease in hemoglobin (Hemoglobin Max ≥ -0.75) was identified as a major risk factor, highlighting the significant clinical importance of maintaining appropriate hemoglobin levels in patients with SALI to improve oxygen delivery and promote liver function recovery.

Furthermore, heart rate is an important hemodynamic status and systemic stress in critically ill patients. In sepsis, persistent tachycardia reflects sympathetic overactivation, systemic inflammation, and impaired cardiovascular compensation, all of which are associated with increased mortality risk [43,44]. Elevated heart rate increases myocardial oxygen demand, worsens tissue hypoperfusion, and can aggravate hepatic injury in SALI patients. In our SHAP and LIME analyses, Heart Rate Max was identified as a key predictor of mortality, emphasizing its clinical value for early risk stratification and timely hemodynamic management.

Additionally, diastolic blood pressure (DiaBP) reflects peripheral vascular resistance and organ perfusion during cardiac diastole. In sepsis, low DiaBP often results from excessive vasodilation, endothelial dysfunction, and reduced vascular tone [45]. Decreased DiaBP impairs hepatic and renal perfusion, exacerbates ischemia, and promotes the progression of sepsis-associated liver injury. Several studies have shown that low diastolic blood pressure is independently associated with increased mortality in sepsis, as it reflects tissue hypoxia and circulatory failure [46,47]. In our SHAP and LIME analyses, a minimum diastolic blood pressure $\geq$ 0.65 was identified as a protective factor, whereas higher thresholds were associated with increased mortality risk. This suggests that insufficient DiaBP may exacerbate liver injury due to hypoperfusion, while maintaining appropriate diastolic pressure helps restore hepatic and systemic function. Therefore, in clinical practice, individualized hemodynamic

management strategies to optimize diastolic blood pressure are crucial for preventing further organ dysfunction in patients with sepsis-associated liver injury.

Overall, these five indicators provide a clinically significant basis for early risk stratification in patients with sepsis-associated liver injury, which not only improves the accuracy of predictions but also offers valuable reference information for evidence-based decision-making in critical care.

4.4. Limitations

Several important limitations must be acknowledged when interpreting the results of this machine learning approach for predicting SALI mortality. First, although features were selected based on statistical methods, existing literature, and expert opinions to ensure the clinical relevance of the model, this process may have introduced preconceived bias, potentially leading to the omission of important predictive variables.

Second, a further limitation of this study is the relatively small sample size, particularly in the external validation cohort. Although the total sample size was 1157 patients, dividing the data into training, internal validation, and two external validation sets resulted in a corresponding reduction in the effective sample size for each analysis. Despite the use of cross-validation and regularization techniques, the limited sample size increases the risk of overfitting and may reduce the stability of the feature importance rankings. In particular, the small sample size in the eICU cohort (N = 106) may affect the stability and reliability of the model performance evaluation, thus limiting the generalizability of the study findings. Given the relatively small size of the development dataset, the observed feature importance patterns may be partially influenced by overfitting.

Third, the performance differences between the MIMIC-III and eICU datasets highlight the challenges of model generalization. Although both the MIMIC-III and eICU databases were employed for external validation, the model demonstrated varying performance across cohorts. It performed well in the MIMIC-III cohort but showed reduced predictive ability in the eICU dataset, reflecting the challenges of applying predictive models to heterogeneous populations and multi-center data sources. Notably, the XGBoost model achieved an accuracy of approximately 60% in the larger and more heterogeneous eICU-CRD cohort. This performance difference highlights a clinically significant gap in the model's generalization ability, emphasizing the challenges of applying models trained on single-center datasets to different multi-center intensive care environments. This performance variability suggests that the relative importance and predictive strength of five key features, including Lactate Min, BUN Min, Heart Rate Max, Hemoglobin Max, and Diastolic Blood Pressure Min, may vary depending on the patient populations and clinical settings. While the strong performance in the MIMIC-III cohort is encouraging, the reliance on data from a single-center database may limit the generalizability of the findings to other healthcare institutions. Such geographic and institutional homogeneity may affect the applicability of the results to broader patient populations or medical centers with different care practices. Furthermore, differences in patient information, clinical practices, and data collection methods between the MIMIC-IV and eICU databases may also have influenced model performance.

In addition, the retrospective study design precludes causal inference, and the interpretability methods applied (SHAP and LIME) can only provide approximate estimates of feature contributions. Therefore, these identified strong predictors should be understood as factors associated with the risk of death, rather than definitive causal factors that determine the risk of death. Their relative importance may differ when applied to populations with different demographic characteristics, disease severity, or treatment strategies.

Finally, the complexity of the XGBoost model may present challenges for clinicians without technical expertise and for resource-limited healthcare settings. The decreased

performance of the pre-trained XGBoost model on the larger, more heterogeneous eICU-CRD dataset further suggests that models trained on relatively small, single-center datasets may not generalize well to external populations without recalibration or retraining. Future research should focus on prospective validation in larger, multicenter cohorts and explore strategies for integrating such models into clinical workflows to enhance decision support in the ICU. Future studies should also assess the stability of key predictive predictors and consider recalibrating or retraining the model when applying it to new populations.

5. Conclusions

This study successfully developed and validated 8 machine learning algorithms for predicting 28-day mortality in patients with Sepsis-Associated Liver Injury (SALI). The investigation leveraged comprehensive clinical data from the MIMIC-IV, MIMIC-III, and eICU databases, encompassing 1192 patients who satisfied stringent SALI diagnostic criteria. Among the eight evaluated algorithms, XGBoost demonstrated the highest overall performance in the internal validation, achieving an AUROC of 0.8556 (95% CI: 0.807–0.898), an accuracy of 0.7702, a recall of 0.8469, and a specificity of 0.7200. SHAP and LIME analyses provided clinically interpretable explanations of feature contributions, highlighting key predictors such as lactate, blood urea nitrogen, hemoglobin, heart rate, and diastolic blood pressure. External validation further emphasized the importance of assessing the model's generalizability across different cohorts. Although the XGBoost model is internally robust, its generalizability to diverse, multi-center clinical settings remains limited and unproven. Future work should prioritize prospective validation in larger, multicenter cohorts to address clinical and institutional heterogeneity, while assessing the stability of key predictors and exploring model recalibration strategies to ensure reliability across diverse populations before widespread clinical application. Despite these limitations, the study indicates that the potential of interpretable machine learning approaches to support early mortality risk stratification in patients with sepsis-associated liver injury (SALI) and to provide transparent, evidence-based support for decision-making in the ICU.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L.; Software, Y.L.; Data curation, Y.L.; Formal analysis, Y.L.; Validation, Y.L.; Visualization, Y.L.; Investigation, Y.L.; Writing—original draft preparation, Y.L.; Writing—review and editing, Y.L. and J.F.; Project administration, Y.L.; Resources, M.P.; Supervision, K.A. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The creation of the database and data sharing were approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT), Beth Israel Deaconess Medical Center (BIDMC), and Philips Healthcare. As all personal health information was de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) regulations, the study was exempt from additional institutional review board (IRB) approval and informed consent requirements.

Informed Consent Statement: Patient consent was waived because all data were fully de-identified and publicly available through PhysioNet.

Data Availability Statement: All datasets analyzed during this study are publicly available from the MIMIC-IV at https://physionet.org/content/mimiciv/2.2/(accessed on 22 March 2025), MIMIC-III at https://physionet.org/content/mimiciii/1.4/(accessed on 10 April 2025), and eICU-CRD at https://physionet.org/content/eicu-crd/2.0/(accessed on 19 May 2025) databases through PhysioNet under approved credentialed access.

Acknowledgments: The authors gratefully acknowledge the valuable guidance and constructive feedback from the faculty members of the Department of Industrial and Systems Engineering at the University of Southern California. Their insightful comments and assistance in language editing and proofreading contributed greatly to the clarity and quality of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Singer, M.; Deutschman, C.S.; Seymour, C.W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G.R.; Chiche, J.D.; Coopersmith, C.M.; et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **2016**, *315*, 801–810. [CrossRef] [PubMed]
2. Jarczak, D.; Kluge, S.; Nierhaus, A. Sepsis—Pathophysiology and therapeutic concepts. *Front. Med.* **2021**, *8*, 628302. [CrossRef]
3. Rudd, K.E.; Johnson, S.C.; Agesa, K.M.; Shackelford, K.A.; Tsoi, D.; Kievlan, D.R.; Colombara, D.V.; Ikuta, K.S.; Kissoon, N.; Finfer, S.; et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *Lancet* **2020**, *395*, 200–211. [CrossRef]
4. Iwashyna, T.J.; Cooke, C.R.; Wunsch, H.; Kahn, J.M. Population burden of long-term survivorship after severe sepsis in older Americans. *J. Am. Geriatr. Soc.* **2012**, *60*, 1070–1077. [CrossRef] [PubMed]
5. Kubes, P.; Jenne, C. Immune responses in the liver. *Annu. Rev. Immunol.* **2018**, *36*, 247–277. [CrossRef] [PubMed]
6. Solhi, R.; Lotfinia, M.; Gramignoli, R.; Najimi, M.; Vosough, M. Metabolic hallmarks of liver regeneration. *Trends Endocrinol. Metab.* **2021**, *32*, 731–745. [CrossRef]
7. Yan, J.; Li, S.; Li, S. The role of the liver in sepsis. *Int. Rev. Immunol.* **2014**, *33*, 498–510. [CrossRef]
8. Chen, J.W.; Liu, C.Y.; Li, S.; Wu, S.W.; Cai, C.; Lu, M.Q. Sepsis-associated liver injury: Mechanisms and potential therapeutic targets. *World J. Gastroenterol.* **2024**, *30*, 4518. [CrossRef]
9. Kobashi, H.; Toshimori, J.; Yamamoto, K. Sepsis-associated liver injury: Incidence, classification and the clinical significance. *Hepatol. Res.* **2013**, *43*, 255–266. [CrossRef]
10. Dellinger, R.P.; Levy, M.; Rhodes, A.; Annane, D.; Gerlach, H.; Opal, S.; Sevransky, J.; Sprung, C.; Douglas, I.; Jaeschke, R.; et al. International guidelines for management of severe sepsis and septic shock. *Intensive Care Med.* **2013**, *39*, 165–228. [CrossRef]
11. Richens, J.G.; Lee, C.M.; Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **2020**, *11*, 3923. [CrossRef] [PubMed]
12. Johnson, A.E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **2023**, *10*, 1. [CrossRef] [PubMed]
13. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
14. Pollard, T.J.; Johnson, A.E.; Raffa, J.D.; Celi, L.A.; Mark, R.G.; Badawi, O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **2018**, *5*, 180178. [CrossRef]
15. Wen, C.; Zhang, X.; Li, Y.; Xiao, W.; Hu, Q.; Lei, X.; Xu, T.; Liang, S.; Gao, X.; Zhang, C.; et al. An interpretable machine learning model for predicting 28-day mortality in patients with sepsis-associated liver injury. *PLoS ONE* **2024**, *19*, e0303469. [CrossRef]
16. Teasdale, G.; Maas, A.; Lecky, F.; Manley, G.; Stocchetti, N.; Murray, G. The Glasgow Coma Scale at 40 years: Standing the test of time. *Lancet Neurol.* **2014**, *13*, 844–854. [CrossRef]
17. Cheng, C.Y.; Kung, C.T.; Chen, F.C.; Chiu, I.M.; Lin, C.H.R.; Chu, C.C.; Kung, C.F.; Su, C.M. Machine learning models for predicting in-hospital mortality in patient with sepsis: Analysis of vital sign dynamics. *Front. Med.* **2022**, *9*, 964667. [CrossRef]
18. Rosenblatt, M.; Tejavibulya, L.; Jiang, R.; Noble, S.; Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Commun.* **2024**, *15*, 1829. [CrossRef]
19. O'brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]
20. Midi, H.; Sarkar, S.K.; Rana, S. Collinearity diagnostics of binary logistic regression model. *J. Interdiscip. Math.* **2010**, *13*, 253–267. [CrossRef]
21. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 26. [CrossRef]
22. Fisher, A.; Rudin, C.; Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81. [CrossRef]
23. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013. [CrossRef]
24. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]

25. Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584. [CrossRef] [PubMed]
26. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–10 [CrossRef]
27. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
28. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
29. Yi, X.; Jin, D.; Huang, S.; Xie, Z.; Zheng, M.; Zhou, F.; Jin, Y. Association between lactate-to-albumin ratio and 28-days all-cause mortality in patients with sepsis-associated liver injury: A retrospective cohort study. *BMC Infect. Dis.* **2024**, *24*, 65. [CrossRef]
30. Xie, T.; Xin, Q.; Cao, X.; Chen, R.; Ren, H.; Liu, C.; Zhang, J. Clinical characteristics and construction of a predictive model for patients with sepsis related liver injury. *Clin. Chim. Acta* **2022**, *537*, 80–86. [CrossRef]
31. Xiao, Y.; Chen, Y.; Huang, R.; Jiang, F.; Zhou, J.; Yang, T. Interpretable machine learning in predicting drug-induced liver injury among tuberculosis patients: Model development and validation study. *BMC Med. Res. Methodol.* **2024**, *24*, 92. [CrossRef]
32. Zhong, T.; Zhuang, Z.; Dong, X.; Wong, K.H.; Wong, W.T.; Wang, J.; He, D.; Liu, S. Predicting antituberculosis drug–induced liver injury using an interpretable machine learning method: Model development and validation study. *JMIR Med. Inform.* **2021**, *9*, e29226. [CrossRef]
33. Liu, S.; Yang, T.; Jiang, Q.; Zhang, L.; Shi, X.; Liu, X.; Li, X. Lactate and lactylation in sepsis: A comprehensive review. *J. Inflamm. Res.* **2024**, *17*, 4405–4417. [CrossRef]
34. Kang, H.E.; Park, D.W. Lactate as a biomarker for sepsis prognosis? *Infect. Chemother.* **2016**, *48*, 252. [CrossRef]
35. Scheiner, B.; Lindner, G.; Reiberger, T.; Schneeweiss, B.; Trauner, M.; Zauner, C.; Funk, G.C. Acid-base disorders in liver disease. *J. Hepatol.* **2017**, *67*, 1062–1073. [CrossRef] [PubMed]
36. Drolz, A.; Horvatits, T.; Rutter, K.; Landahl, F.; Roedl, K.; Meersseman, P.; Wilmer, A.; Kluwe, J.; Lohse, A.W.; Kluge, S.; et al. Lactate improves prediction of short-term mortality in critically Ill patients with cirrhosis: A multinational study. *Hepatology* **2019**, *69*, 258–269. [CrossRef] [PubMed]
37. Harazim, M.; Tan, K.; Nalos, M.; Matejovic, M. Blood urea nitrogen-independent marker of mortality in sepsis. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc. Czech. Repub.* **2023**, *167*, 24–29. [CrossRef] [PubMed]
38. Li, X.; Zheng, R.; Zhang, T.; Zeng, Z.; Li, H.; Liu, J. Association between blood urea nitrogen and 30-day mortality in patients with sepsis: A retrospective analysis. *Ann. Palliat. Med.* **2021**, *10*, 11653–11663. [CrossRef]
39. Han, K.; Tao, Y.; Wang, J.; Lu, J. Prognostic value of blood urea nitrogen to albumin ratio in septic patients with acute kidney injury—A retrospective study based on MIMIC database. *Front. Med.* **2025**, *12*, 1510919. [CrossRef]
40. Vincent, J.L.; Baron, J.F.; Reinhart, K.; Gattinoni, L.; Thijs, L.; Webb, A.; Meier-Hellmann, A.; Nollet, G.; Peres-Bota, D.; ABC Investigators. Anemia and blood transfusion in critically ill patients. *Jama* **2002**, *288*, 1499–1507. [CrossRef]
41. Corwin, H.L.; Gettinger, A.; Pearl, R.G.; Fink, M.P.; Levy, M.M.; Abraham, E.; MacIntyre, N.R.; Shabot, M.M.; Duh, M.S.; Shapiro, M.J. The CRIT Study: Anemia and blood transfusion in the critically ill—Current clinical practice in the United States. *Crit. Care Med.* **2004**, *32*, 39–52. [CrossRef]
42. Jung, S.M.; Kim, Y.J.; Ryoo, S.M.; Kim, W.Y. Relationship between low hemoglobin levels and mortality in patients with septic shock. *Acute Crit. Care* **2019**, *34*, 141. [CrossRef]
43. Shen, Y.; Wang, J.; Cao, Q.; Wu, Y.; Wang, Q.; Wang, N.; Shao, M. Maximum heart rate and mortality in sepsis patients: A retrospective cohort study. *Intern. Emerg. Med.* **2025**, *20*, 1–10. [CrossRef]
44. Ning, Y.L.; Li, W.J.; Lu, X.; Zhang, Y.; Zhang, J.W.; Zhou, J.H. Association between heart rate and mortality in patients with septic shock: An analysis revealed by time series data. *BMC Infect. Dis.* **2024**, *24*, 1088. [CrossRef]
45. Cecconi, M.; De Backer, D.; Antonelli, M.; Beale, R.; Bakker, J.; Hofer, C.; Jaeschke, R.; Mebazaa, A.; Pinsky, M.R.; Teboul, J.L.; et al. Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine. *Intensive Care Med.* **2014**, *40*, 1795–1815. [CrossRef]
46. Maheshwari, K.; Nathanson, B.H.; Munson, S.H.; Khangulov, V.; Stevens, M.; Badani, H.; Khanna, A.K.; Sessler, D.I. The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. *Intensive Care Med.* **2018**, *44*, 857–867. [CrossRef]
47. Dünser, M.W.; Ruokonen, E.; Pettilä, V.; Ulmer, H.; Torgersen, C.; Schmittinger, C.A.; Jakob, S.; Takala, J. Association of arterial blood pressure and vasopressor load with septic shock mortality: A post hoc analysis of a multicenter trial. *Crit. Care* **2009**, *13*, R181. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.