Development of machine learning models to predict 28-Day mortality in patients with sepsis-associated liver injury

Yupeng Li, Beulah Kannan, Yexi Zhu

Abstract

Background: Sepsis-Associated Liver Injury (SALI) is an independent risk factor for mortality in sepsis patients, contributing significantly to adverse outcomes in intensive care settings. This comprehensive study aimed to develop and validate interpretable machine learning models to predict 28-day mortality in sepsis-associated liver injury (SALI) patients using extensive data from MIMIC-IV (v2.2) and MIMIC-III (v1.4) databases. The study cohort comprised 1,192 patients (834 from MIMIC-IV and 358 from MIMIC-III), with MIMIC-IV data randomly divided into training (70%) and internal validation (30%) sets, while MIMIC-III served as external validation to assess model generalizability across different patient populations.

Methods: A rigorous data preprocessing workflow was implemented, where features with greater than 20% missing values were removed, and remaining missing data were addressed through multiple interpolation techniques. Permutation importance was employed for feature selection, resulting in the identification of 30 clinically relevant parameters from an initial pool of 105 features after aggregation functions were applied. The study developed and compared eight different machine learning models: Random Forest (RF), Logistic Regression, Decision Tree, Extreme Gradient Boost (XGBoost), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Generalized Linear Models with cross-validation (CV_glmnet), and Linear Discriminant Analysis (LDA).

Results: In the internal validation cohort, Support Vector Machine (SVM) demonstrated superior performance with the highest Area Under the Curve (AUC) of 0.9159 (95% CI: 0.877-0.948), followed closely by XGBoost with an AUC of 0.9127 (95% CI: 0.877-0.948). For external validation, RF achieved the highest AUC of 0.9317 (95% CI: 0.905-0.957), indicating excellent discriminative ability and generalizability. Shapley additive explanation (SHAP) analysis was implemented to enhance model interpretability, identifying key predictive features including heart rate, partial thromboplastin time (PTT), aspartate aminotransferase (AST), lactate levels, and systolic blood pressure. These features align with clinical understanding of SALI pathophysiology, validating the model's biological plausibility. The developed models consistently outperformed traditional disease severity scoring systems such as Glasgow Coma Scale (GCS) score, Sequential Organ Failure Assessment (SOFA) score, quick SOFA (qSOFA) score, and Acute Physiology and Chronic Health Evaluation II (APACHE II) score.

Conclusion: The Support Vector Machine (SVM) algorithm produced a highly accurate and interpretable model for predicting mortality in patients with sepsis-associated liver injury (SALI) across diverse populations. Despite limitations, including reliance on single-center datasets, the models demonstrated robust predictive capabilities and strong generalizability, offering significant potential for early identification of high-risk SALI patients.

Keywords: Sepsis, Machine Learning, Liver Injury, SALI, MIMIC-IV, MIMIC-III, Predictive modeling, External Validation

1. Introduction

Sepsis remains one of the most challenging conditions in critical care medicine, representing a dysregulated host response to infection that leads to life-threatening organ dysfunction [3]. This

complex syndrome continues to pose significant challenges to healthcare systems worldwide, with recent epidemiological studies demonstrating concerning trends in both incidence and mortality [6]. The economic implications are equally substantial, as healthcare expenditures for sepsis management consume a disproportionate share of hospital resources compared to other conditions requiring inpatient care [7].

The liver occupies a central position in the body's defense against infection, serving as both a critical immune organ and metabolic regulator. This vital organ contributes to host protection through multiple mechanisms including pathogen clearance, toxin neutralization, and modulation of inflammatory responses [9]. During sepsis, hepatic function becomes compromised due to various insults including microcirculatory dysfunction, inflammatory mediator release, and cellular energetic failure [10].

Sepsis-Associated Liver Injury (SALI) represents a significant complication that develops in a substantial proportion of septic patients. The pathophysiology involves complex interactions between direct pathogen effects, microcirculatory disturbances, inflammatory cascades, and metabolic derangements [11, 12]. Clinically, SALI manifests as either cholestatic dysfunction or hypoxic hepatitis, with diagnostic criteria typically centered on elevations in total bilirubin and coagulation parameters [9]. Traditional severity assessment tools like the Sequential Organ Failure Assessment (SOFA) incorporate hepatic parameters but lack specificity for SALI outcomes [10].

Epidemiological research indicates that SALI affects a significant portion of sepsis patients and substantially increases mortality risk [13]. The development of liver dysfunction in the setting of sepsis has been associated with nearly doubled mortality rates in some studies, highlighting the

critical need for improved prognostic tools [4]. Despite this clear association with adverse outcomes, current clinical practice lacks robust early warning systems specifically designed to identify patients at highest risk for SALI-related mortality.

Machine learning approaches offer promising alternatives to traditional statistical methods for outcome prediction in critically ill populations [14]. By leveraging complex algorithms capable of identifying non-linear relationships and interactions between variables, these techniques may overcome limitations of conventional scoring systems [15]. Recent studies have demonstrated the potential of machine learning to enhance predictive accuracy across various critical care scenarios, though applications specific to SALI remain limited.

This study aims to develop and validate an interpretable machine learning model for early prediction of 28-day mortality in patients with SALI. By utilizing comprehensive clinical data from established critical care databases and employing advanced analytical techniques, this study seeks to create a tool that provides clinicians with actionable information to guide therapeutic interventions. The ultimate goal is to facilitate earlier recognition of high-risk patients, enable more timely and targeted treatments, and potentially improve survival outcomes in this vulnerable population.

2. Methods

2.1. Data Source

This study extracted data from two large, publicly available, and comprehensive clinical care databases, the Medical Information Mart for Intensive Care IV (MIMIC-IV, v2.2) [1] from 2008 to 2019 and the Medical Information Mart for Intensive Care III (MIMIC-III, v1.4) [2] from 2001 to 2012. Both databases contain detailed patient information, including 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).

2.2. Study Population

The study cohort consisted of adult patients (Age ≥ 18) diagnosed with sepsis-associated liver injury (SALI) who had an intensive care unit (ICU) stay of at least 24 hours in the MIMIC-IV database. Sepsis was defined based on the Sepsis-3 criteria [3] as suspected infections with a Sequential Organ Failure Assessment (SOFA) score ≥ 2. Sepsis-associated liver injury (SALI) was 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. [4] 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 hours of ICU admission. The same inclusion and exclusion criteria were applied to the MIMIC-III databases for external validation. Finally, 834 patients from the MIMIC-IV database and 358 patients from the MIMIC-III database were included in the final study cohort. The detailed patient selection process is illustrated in Figure 1.

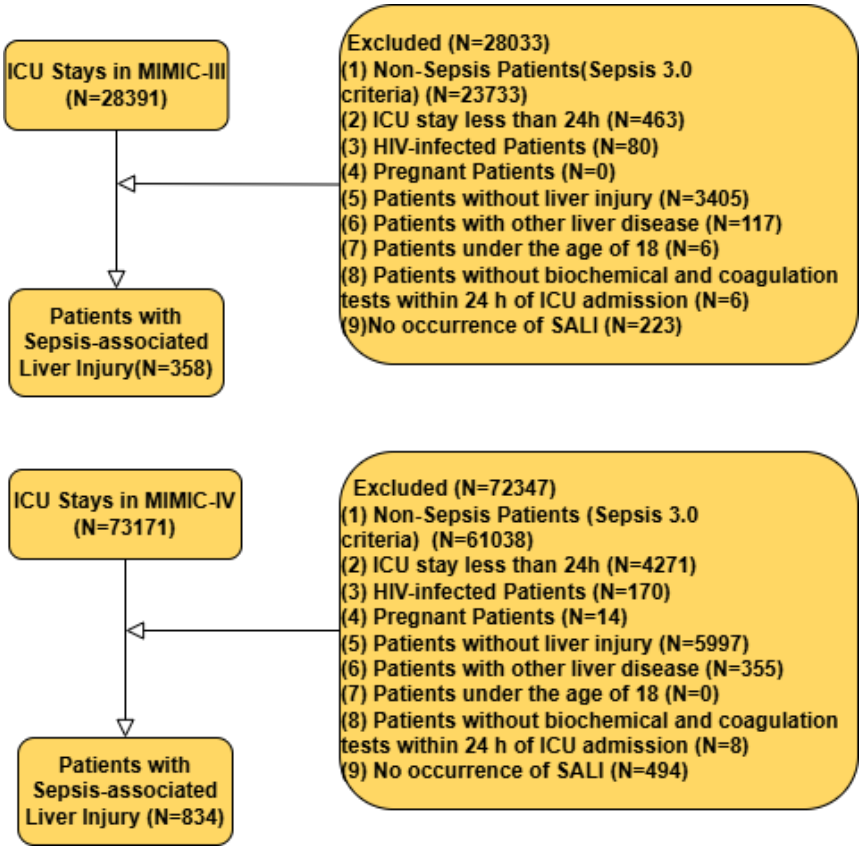

Figure 1: The flow chart of patient selection process

2.3. Data Collection

The data within the first 24 hours of ICU admission were extracted by Structured Query Language(SQL), MySQL, and Python, based on existing literature [5], clinical relevance, and expert recommendations to capture early clinical status. All diagnoses were identified based on the International Classification of Diseases, Tenth Revision, Clinical Modifications (ICD-10-CM) and Ninth Revision, Clinical Modifications (ICD-9-CM). Initially, a total of 69 clinical features were extracted from the MIMIC-IV database, categorized into 8 groups: demographics (e.g. age, ICU length of stay, and gender), vital signs (e.g., heart rate, respiratory rate, and temperature), laboratory indicators (e.g., partial thromboplastin time (PTT), total bilirubin, and blood glucose), comorbidities (e.g., hypertension, congestive heart failure, and

diabetes), infection sites (e.g., intestinal infection, urinary infection, and lung infection), interventions (e.g., vasopressor use, mechanical ventilation use, and antibiotic use), clinical measurements (e.g., total urine output on day 1 and glomerular filtration rate (GFR)) and severity scores (e.g., Glasgow Coma Scale (GCS) score, Sequential Organ Failure Assessment (SOFA) score, and quick SOFA (qSOFA) score). For vital signs and laboratory indicators, the minimum and maximum values during the first 24 hours of ICU admission were computed to characterize physiological and biochemical fluctuations. According to the Surviving Sepsis Campaign (SSC) International Guidelines [4], this study focused on 28-day mortality as a primary outcome for evaluating short-term survival in patients with sepsis. The same set of features was retrieved from the MIMIC-III database to ensure consistency between study cohorts and to minimize potential bias related to feature selection. The complete list of 69 clinical features is provided in Table 1.

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 (DSP) 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 Antibiotic 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

2.4. Ethics Statement

The databases for this study were approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC). All personal health information was de-identified by strict evaluation standards in compliance with the Health Insurance Portability and Accountability Act (HIPAA) regulations, and the requirement for individual informed consent was waived. [1, 2]

2.5. Data preprocessing

Preprocessing was performed consistently for both internal validation (MIMIC-IV) and external validation (MIMIC-III) cohorts. The datasets were initially divided into training and test sets. Missing values were handled independently within each dataset to prevent data leakage. Features with more than 20% missing values were removed based only on the training set. For the remaining features, missing values were imputed using the median value computed from the training set and subsequently applied to the test set.

Categorical variables were preprocessed based on their nature: binary categorical variables were label encoded, while multiclass categorical variables were one-hot encoded. For high-cardinality categorical variables, the most frequent categories were retained, and infrequent categories were grouped into an 'Other' category to prevent sparsity.

Outliers were addressed through winsorization by capping extreme values. Variables exhibiting high positive skewness were log-transformed to approximate normality. Features with near-zero

variance were excluded to improve model stability. Additionally, highly correlated features (correlation coefficient > 0.85) were removed to minimize multicollinearity, and variance inflation factor (VIF) analysis was performed to further refine the feature set. Finally, all features were scaled and standardized prior to model training. The flow chart of data preprocessing is provided in Figure 2.

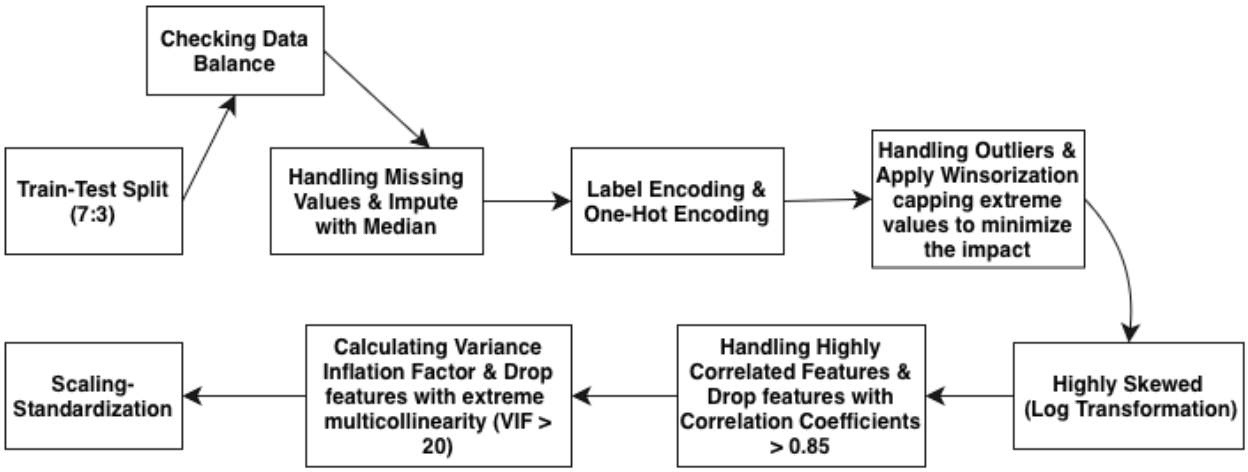

Figure 2: The flow chart of data preprocessing

2.6. Feature selection

Permutation importance was applied to the internal validation set to identify the most influential features contributing to model performance. Permutation importance measures the decrease in model performance when the values of a single feature are randomly shuffled, thereby disrupting the relationship between that feature and the outcome. A greater decrease in the ROC-AUC score indicates higher feature importance.

In this study, the internal dataset was randomly divided into training and test sets. Permutation importance [17] was computed on the held-out internal test set, using 10 repeated shuffles per feature to ensure stability of the estimates. Based on the results, the top 30 features were selected

for further model refinement. The most important features identified included minimum heart rate, minimum PTT, minimum lactate levels, maximum diastolic blood pressure, total urine output, ICU length of stay, minimum aspartate aminotransferase levels, minimum SpO_2, minimum diastolic blood pressure, maximum fibrinogen levels, minimum blood urea nitrogen (BUN), maximum hematocrit levels, minimum total bilirubin, minimum platelet count, maximum systolic blood pressure, maximum glucose serum levels, age, maximum white blood cell count, maximum total bilirubin, minimum albumin, presence of diabetes, minimum serum calcium, minimum systolic blood pressure, minimum hematocrit levels, maximum lactate levels, SOFA score, maximum albumin, minimum serum potassium, minimum direct bilirubin, and maximum platelet count. These features were used to retrain the models, improving both predictive performance and interpretability. The 30 selected features with mean importance are shown in Figure 3.

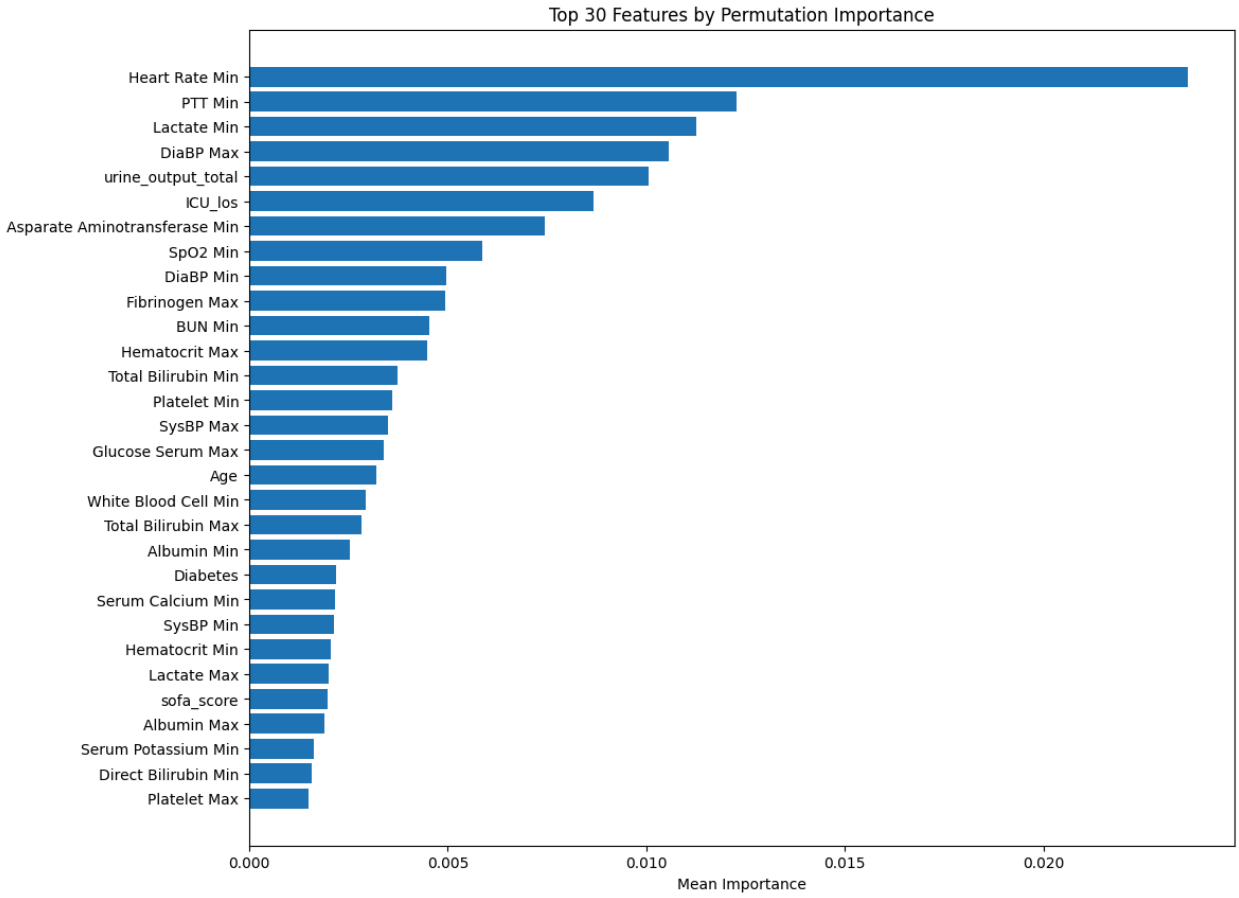

Figure 3: 30 selected features with mean importance

2.7. Statistical analysis

Continuous variables were summarized as medians with interquartile ranges (IQR), and categorical variables as counts with percentages. The rank-sum test was used for continuous variables, and the Chi-square test for categorical variables. After data preprocessing and feature selection, eight commonly used machine learning models were developed to predict 28-day mortality in patients with sepsis-related liver injury. Model performance was assessed using AUC, accuracy, precision, recall, and specificity. The best-performing model was interpreted using SHAP (Shapley Additive Explanations) values. All statistical analyses were conducted using Python. Two-sided tests were used, and P-values < 0.05 were considered statistically significant.

2.8. Modeling

Multiple machine learning models were developed and evaluated to predict 28-day mortality among patients. The following classifiers were implemented: logistic regression, logistic regression with ElasticNet regularization (via cross-validation), random forest, support vector machine (SVM), k-nearest neighbors (KNN), decision tree, linear discriminant analysis (LDA), and extreme gradient boosting (XGBoost).

For each model, a hyperparameter tuning strategy was applied using GridSearchCV with 10-fold cross-validation, optimizing for the area under the receiver operating characteristic curve (ROC-AUC) on the internal training set. Hyperparameter grids were designed specific to each algorithm, including parameters such as regularization strength and mixing ratios for logistic regression, the number of estimators and maximum depth for random forests and XGBoost, kernel types for SVMs, and neighbor counts for KNNs.

Following hyperparameter optimization, the models were trained using only the top 30 features identified through permutation importance. Model performance was assessed on the internal test set using several metrics, including ROC-AUC, accuracy, precision, recall, F1-score, and specificity. A custom probability threshold of 0.3 was applied to convert predicted probabilities into class labels, enhancing sensitivity to the minority class.

Bootstrapping was performed to estimate 95% confidence intervals for the ROC-AUC scores, providing robust statistical evaluation of model performance. Among the tested models, the support vector machine and XGBoost classifiers demonstrated the highest predictive performance.

3. Results

3.1. Baseline characteristics

A total of 834 patients with SALI were analyzed, comprising 505 survivors and 329 non-survivors. The median age was similar between groups, though non-survivors were slightly older (65.2 vs. 63.3 years, $p = 0.0594$). Gender distribution was comparable (62% male overall, $p=0.8326$). Racial distribution differed significantly, with fewer non-survivors identifying as White and more as "Other" ($p = 0.0253$ and $p = 0.0002$, respectively). Non-survivors had higher BMI (38.8 vs. 28.1, $p = 0.0075$), lower albumin, and elevated markers of severity, including higher SOFA and qSOFA scores ($p < 0.0001$ for both). Comorbidities like congestive heart failure, diabetes, hypertension, and renal disease were significantly more common in non-survivors ($p < 0.01$), while myocardial infarction and chronic pulmonary disease showed no significant differences. Non-survivors had higher lactate, anion gap, creatinine, bilirubin, INR, and liver enzymes, indicating greater metabolic derangement and hepatic dysfunction. Hematologic parameters such as hemoglobin, hematocrit, and red blood cell count were significantly lower among non-survivors ($p< 0.0001$). Non-survivors had shorter ICU stays, lower urine output, and reduced PaO_2/FiO_2 ratios, all indicating more critical illness. Infection patterns also varied: non-survivors had fewer skin, urinary, and intestinal infections but were more likely to have catheter-related infections. Vital signs revealed lower temperature and Oxygen Saturation (SpO_2) and a lower Mean Arterial Pressure (MAP) among non-survivors. Vasopressor use was more frequent among non-survivors (86% vs. 78.2%, $p = 0.0063$). All patients had liver disease, received mechanical ventilation, and antibiotics. Scores reflecting illness severity, such as APACHE II, were significantly higher in non-survivors (30.1 vs. 26.9, $p < 0.0001$), as were total bilirubin and Asparate aminotransferase (AST) levels. Renal function, measured by Glomerular Filtration Rate (GFR), was also lower in non-survivors (25.2 vs. 32.3, p

= 0.0016). The baseline characteristics of the entire cohort with survival and non-survival groups from the MIMIC-IV database are shown in S1 Table.

3.2. Model performance

In our internal validation using MIMIC-IV, the Support Vector Classifier model outperformed other classifiers with an AUC of **0.9159 (**95% CI: [0.877, 0.948]), accuracy of **0.796**, precision of **0.693**, recall of **0.868**, and specificity of **0.75**, demonstrating strong predictive ability. These findings are consistent with the benchmark study by Wen et al. (2024) [5], where Random Forest achieved the highest AUC (**0.79**), along with comparable accuracy (**0.746**) and specificity (**0.593**). Most other models, including Logistic Regression, SVM, LDA, KNN, XGBoost, Decision Trees, Logistic Regression with Elastic Net Regularization achieved AUC scores around **0.78–0.91**, indicating better and competitive performance compared to the performances by Wen et al. (2024) [5]. The model performance for internal validation is provided in Table 3. Our ROC curve analysis further supports this, as the Support Vector classifier model consistently showed a superior true positive rate across various thresholds compared to other models. Other classifiers such as XGBoost, Random Forest, and Logistic Regression performed closely, yet with slightly lower AUC values and F1 scores. The ROC curve for internal validation is illustrated in Figure 5.

Table 3: Model performance for internal validation

Models	AUC	AUC 95% CI	Accuracy	Precision	Recall	F1-Score	Specificity
SVM	**0.9159**	**[0.877, 0.948]**	**0.7968**	**0.6935**	**0.8687**	**0.7713**	**0.7500**
XGBoost	0.9127	[0.877, 0.948]	0.8207	0.7647	0.7879	0.7761	0.8421
CV_Glmnet	0.8898	[0.843, 0.931]	0.7729	0.6567	0.8889	0.7554	0.6974
Random Forest	0.8874	[0.841, 0.926]	0.7410	0.6232	0.8687	0.7554	0.6579
Logistic Regression	0.8852	[0.836, 0.928]	0.7610	0.6403	0.8990	0.7479	0.6711
LDA	0.8808	[0.837, 0.923]	0.7888	0.7018	0.8081	0.7512	0.7763
KNN	0.8662	[0.819, 0.909]	0.7888	0.7500	0.7225	0.7225	0.8487
Decision Tree	0.7856	[0.725, 0.839]	0.7410	0.6491	0.7475	0.6948	0.7368

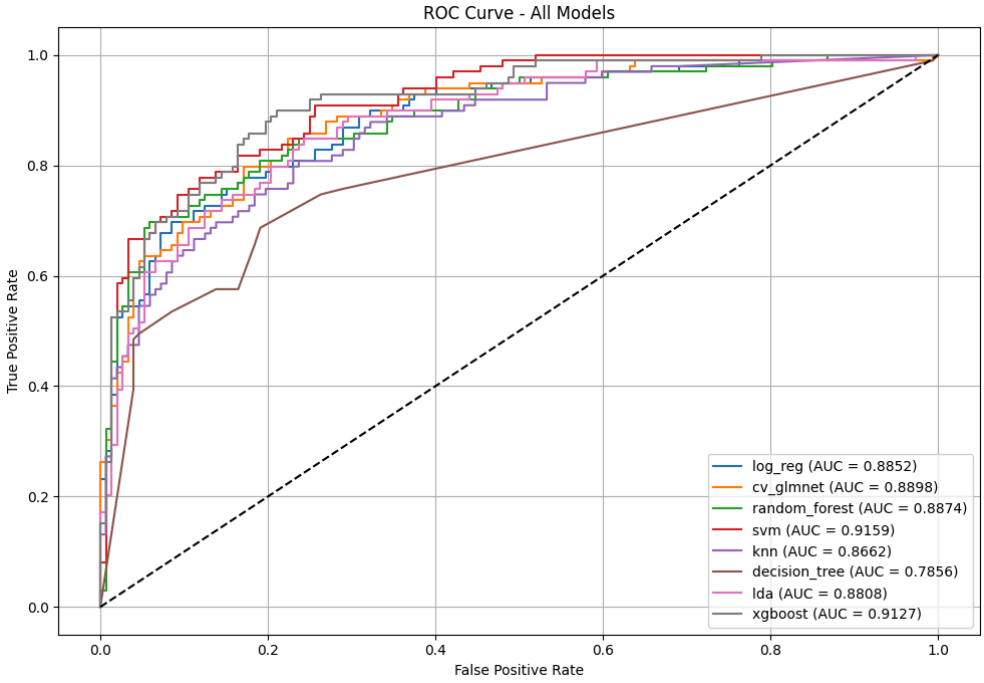

Figure 5: ROC curves for internal validation

During external validation on MIMIC-III data, our models exhibited stronger performance compared to those reported in the reference study. In particular, **Random Forest** achieved an AUC of **0.9317** (95% CI: [0.905, 0.957]), recall of **0.984**, and F1-score of **0.6949**, outperforming Wen et al. (2024)'s external validation AUC of **0.77** for the same model. Additionally, our **XGBoost** and **SVM** models showed competitive AUCs of **0.9191**, with XGBoost delivering the highest F1-score (**0.7712**), suggesting improved generalization and better class balance handling. The model performance for external validation is shown in Table 4. In comparison, the Wen et al. study reported generally lower AUC values (e.g., Logistic Regression: 0.74, CV_Glmnet: 0.75), and lower specificity scores across all models. This discrepancy may be attributed to differences in preprocessing, feature engineering, or sample characteristics. The ROC curves from our evaluation displayed a clear separation from the diagonal baseline, with tighter clustering among top-performing models, affirming robust discriminatory power. The ROC curve for external validation is illustrated in Figure 6.

Table 4: Model performance for external validation

Models	AUC	AUC 95% CI	Accuracy	Precision	Recall	F1-Score	Specificity
Random Forest	**0.9317**	**[0.905, 0.957]**	**0.6983**	**0.5371**	**0.9840**	**0.6949**	**0.5451**
SVM	0.9191	[0.888, 0.947]	0.7095	0.5493	0.9360	0.6923	0.588
XGBoost	0.9191	[0.886, 0.948]	0.8045	0.6519	0.9440	0.7712	0.7296
Logistic Regression	0.9015	[0.868, 0.933]	0.6816	0.5236	0.9760	0.6816	0.5236
CV_Glmnet	0.894	[0.861, 0.927]	0.676	0.5198	0.9440	0.6705	0.5322
KNN	0.885	[0.845, 0.918]	0.7598	0.6102	0.8640	0.7152	0.7039
LDA	0.8821	[0.847, 0.918]	0.7011	0.5441	0.8880	0.6748	0.6009
Decision Tree	0.8138	[0.766, 0.86]	0.7458	0.5988	0.8240	0.6936	0.7039

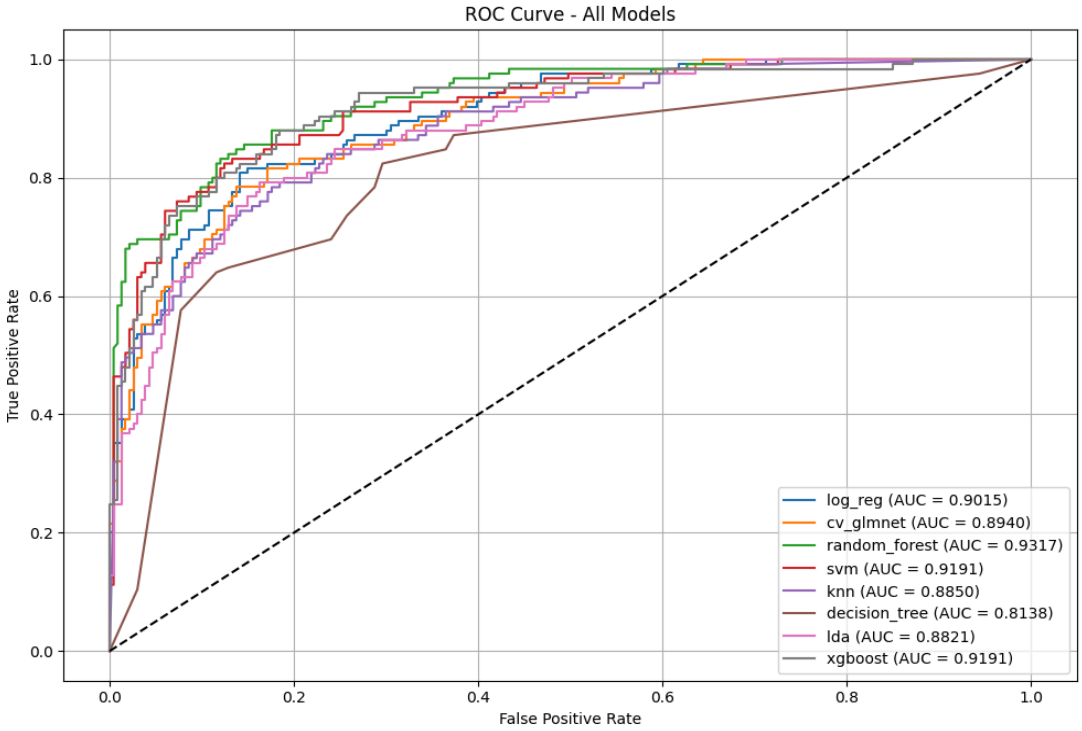

Figure 6: ROC curves for external validation

3.3. SHAP Value Analysis

To enhance model interpretability and understand the contribution of individual features to the predicted risk of 28-day mortality, SHapley Additive exPlanations (SHAP) analysis was performed on the best-performing XGBoost model. SHAP values provide a consistent and theoretically grounded approach to measure each feature's impact on the model's output for individual predictions. [16]

A SHAP summary plot was generated to visualize the distribution of feature impacts across the dataset. Features with higher mean absolute SHAP values were considered more influential. The top contributors included minimum heart rate, minimum PTT, minimum aspartate aminotransferase levels, minimum lactate levels, minimum systolic blood pressure, and total urine output. In the SHAP value plot, higher values of features such as minimum heart rate and minimum PTT were associated with increased predicted mortality risk, while lower values were generally protective.

Additionally, a bar plot of mean absolute SHAP values was constructed to rank the features based on their overall importance. This analysis confirmed the clinical relevance of key predictors and validated the results obtained through permutation importance, providing deeper insights into the model's decision-making process. The SHAP value plot and bar plot are illustrated in Figure 7A and Figure 7B, respectively.

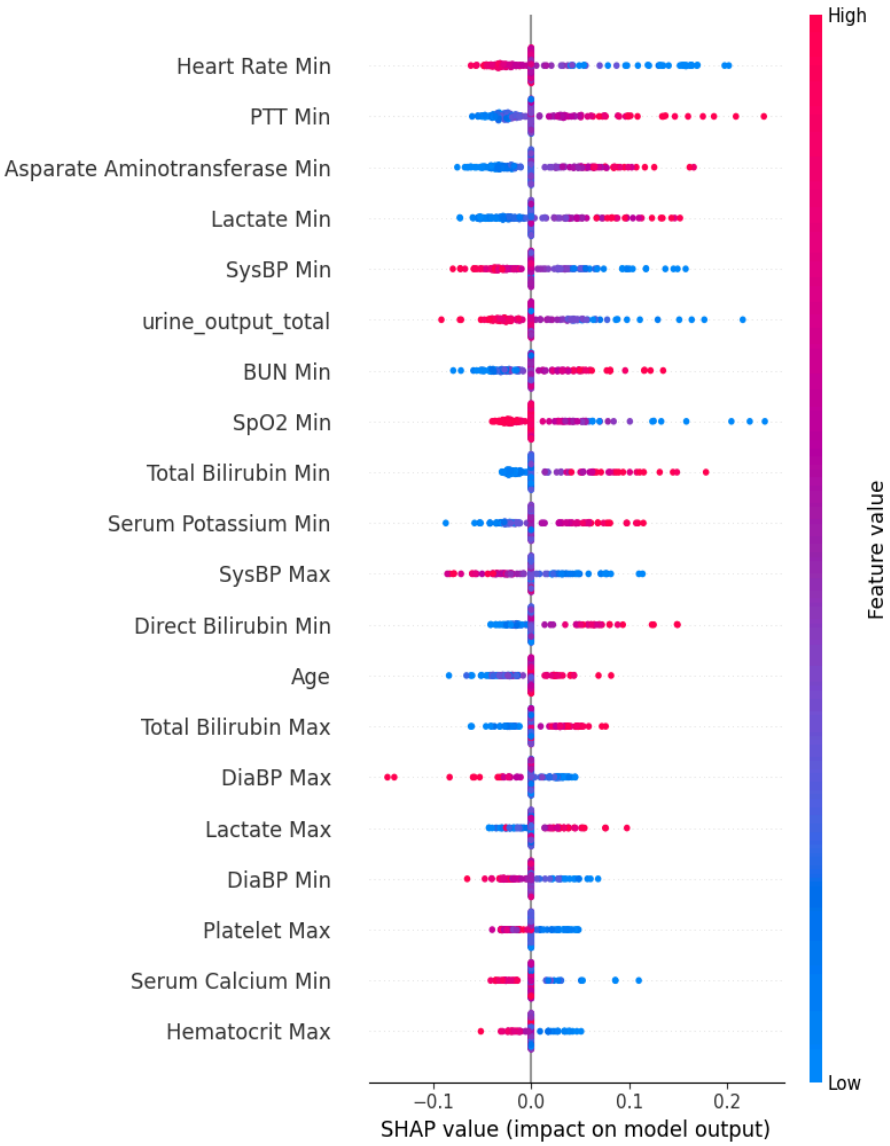

Figure 7: SHAP summary chart.
Feature Impact on SVM Predictions (SHAP Values).

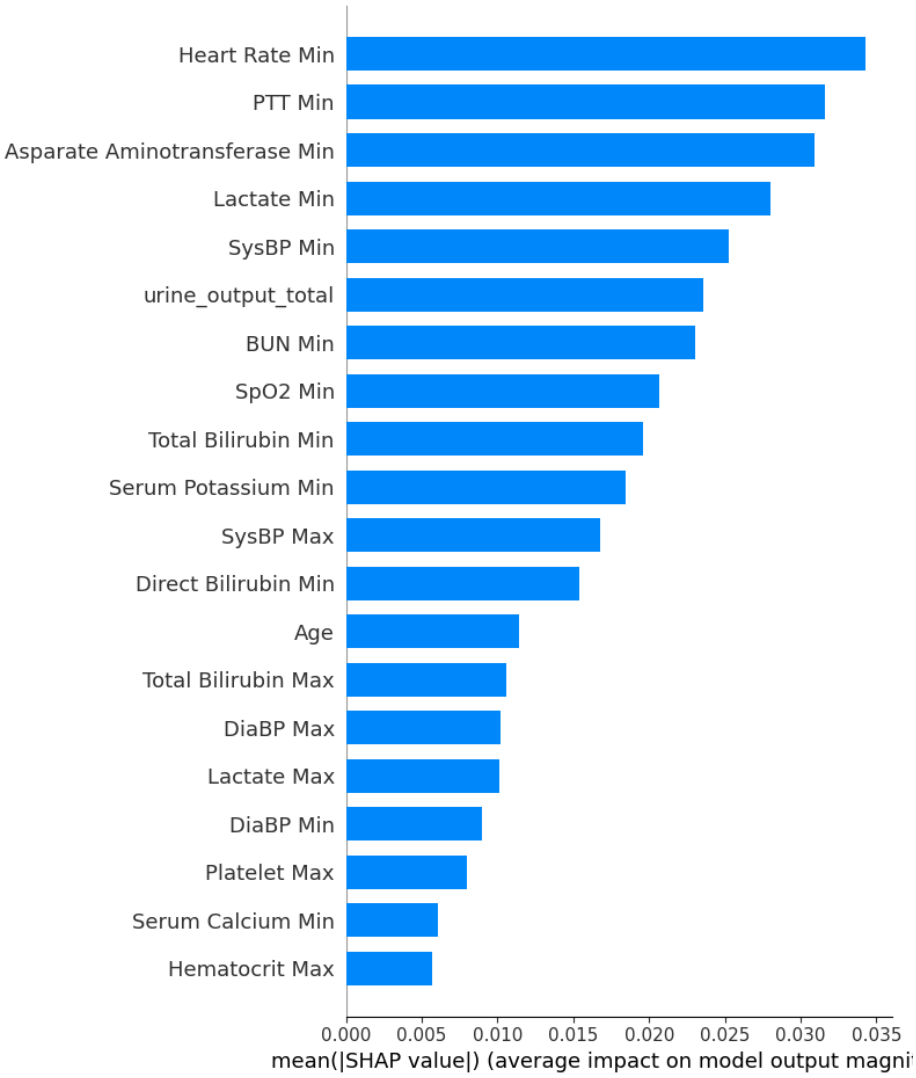

Figure 7: SHAP summary chart.
A. Mean absolute SHAP values for each feature.

4. Discussion

Several important limitations must be acknowledged when interpreting the results of this machine learning approach for predicting SALI mortality. The study's reliance on data from a single healthcare system database represents a significant methodological limitation. The MIMIC databases contain rich clinical information but are based on data from one hospital, which may limit the effectiveness of the results in other settings. This geographic and institutional

homogeneity may impact the generalizability of findings to more diverse patient populations or healthcare settings with different practice patterns. The utilization of the same database source for both model development and validation introduces the potential for overfitting. Despite the rigorous cross-validation approach and the temporal separation between MIMIC-III and MIMIC-IV cohorts, models may become excessively tailored to the particular characteristics, coding practices, and treatment protocols of this specific institution. This institutional bias could limit the algorithm's effectiveness when applied to external datasets from different healthcare systems with varying patient characteristics, clinical workflows, or documentation methods. The demographic composition of the study population warrants careful consideration. The MIMIC databases, while extensive, may not adequately represent the full diversity of sepsis presentations across different racial, ethnic, socioeconomic, and geographic contexts. Specific subpopulations might be underrepresented, potentially limiting the model's performance in certain patient groups. The transferability of the predictive algorithms to more heterogeneous populations requires further validation in multi-center studies encompassing broader demographic profiles. Moreover, the inclusion and exclusion criteria employed in defining SALI may have introduced selection bias, potentially omitting certain patient phenotypes that do not meet conventional diagnostic thresholds but nonetheless experience clinically significant hepatic dysfunction during sepsis. The retrospective nature of the data collection process further introduces the possibility of documentation inconsistencies that could affect variable integrity.

5. Conclusion

This research successfully developed and validated multiple 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-III and MIMIC-IV

databases, encompassing 1,192 patients who satisfied stringent SALI diagnostic criteria. Systematic feature engineering and selection methodologies identified critical mortality predictors. SHAP value analysis revealed heart rate minimum, partial thromboplastin time minimum, aspartate aminotransferase minimum, and lactate minimum as the most significant prognostic indicators. These findings corroborate established clinical understanding of sepsis pathophysiology, where hemodynamic parameters and hepatic dysfunction markers significantly influence patient outcomes. Comparative performance evaluation of eight machine learning algorithms demonstrated that Support Vector Machine (SVM) yielded superior results in internal validation, with area under the receiver operating characteristic curve (AUC) values of 0.9159. These metrics substantially exceed previously reported benchmarks from Wen et al.'s 2024 [5] study, which documented a maximum AUC of 0.79 for Random Forest implementations. External validation confirmed the Random Forest model's robust generalizability, achieving an AUC of 0.9317 with new patient cohorts.

Supporting information

S1 Table. Baseline Characteristics of the study cohort from MIMIC-IV (CSV)

https://drive.google.com/file/d/1TQPhbgx6jb3KqtSdITrjoAyhG16XWZ5s/view?usp=sharing

References

[1] Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LW. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data. 2023 Jan 3;10(1):1.

[2] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.

[3] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016 Feb 23;315(8):801-10.

[4] Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal S, Sevransky J, Sprung C, Douglas I, Jaeschke R. International guidelines for management of severe sepsis and septic shock. Intensive Care Med. 2013;39(2):165-228.

[5] Wen C, Zhang X, Li Y, Xiao W, Hu Q, Lei X, Xu T, Liang S, Gao X, Zhang C, Yu Z. An interpretable machine learning model for predicting 28-day mortality in patients with sepsis-associated liver injury. Plos one. 2024 May 20;19(5):e0303469.

[6] Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, Colombara DV, Ikuta KS, Kissoon N, Finfer S, Fleischmann-Struzek C. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet. 2020 Jan 18;395(10219):200-11.

[7] Shankar-Hari M, Saha R, Wilson J, Prescott HC, Harrison D, Rowan K, Rubenfeld GD, Adhikari NK. Rate and risk factors for rehospitalisation in sepsis survivors: systematic review and meta-analysis. Intensive care medicine. 2020 Apr;46:619-36.

[8] Kubes P, Jenne C. Immune responses in the liver. Annual review of immunology. 2018 Apr 26;36(1):247-77.

[9] Zhang X, Liu H, Hashimoto K, Yuan S, Zhang J. The gut–liver axis in sepsis: interaction mechanisms and therapeutic potential. Critical care. 2022 Jul 13;26(1):213.

[10] Yan J, Li S, Li S. The role of the liver in sepsis. International reviews of immunology. 2014 Nov 2;33(6):498-510.

[11] Solhi R, Lotfinia M, Gramignoli R, Najimi M, Vosough M. Metabolic hallmarks of liver regeneration. Trends in Endocrinology & Metabolism. 2021 Sep 1;32(9):731-45.

[12] Lelubre C, Vincent JL. Mechanisms and treatment of organ failure in sepsis. Nature Reviews Nephrology. 2018 Jul;14(7):417-27.

[13] Minemura M, Tajiri K, Shimizu Y. Liver involvement in systemic infection. World journal of hepatology. 2014 Sep 27;6(9):632.

[14] Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, Machado FR, Mcintyre L, Ostermann M, Prescott HC, Schorr C. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Critical care medicine. 2021 Nov 1;49(11):e1063-143.

[15] Chen Q, Zhang L, Ge S, He W, Zeng M. Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study. PeerJ. 2019 Jun 10;7:e7083.

[16] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[17] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.