Eurasian J. Chem. Med. Pet. Res., **(2027),** Volume 6, Issue 2, 218-230

Eurasian Journal of Chemical, Medicinal and Petroleum Research

Journal homepage: https://www.ejcmpr.com/

DOI: https://doi.org/10.5281/zenodo.24564290

Machine Learning-Based Prediction of Sepsis in Intensive Care Units Using Electronic Health Records

Ali Emami

Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

Article info

Received: 10.09.2026
Accepted: 25.11.2026
Available Online: 26.11.2026
Checked for Plagiarism: Yes

Keywords:
sepsis prediction, machine learning, intensive care unit, electronic health records, XGBoost

ABSTRACT

Sepsis remains a leading cause of mortality in intensive care units (ICUs) worldwide, and early identification is critical for improving patient outcomes. Machine learning (ML) models trained on electronic health records (EHR) have emerged as promising tools for sepsis prediction, yet their clinical translation remains limited by heterogeneity in study design, validation maturity, and dataset overlap. This study aimed to develop and evaluate ML models for predicting sepsis onset in ICU patients using the Medical Information Mart for Intensive Care (MIMIC-IV) database. We extracted data from 12,845 ICU admissions and constructed prediction models using logistic regression, random forest, XGBoost, and long short-term memory networks. The primary outcome was sepsis onset within 24 hours of ICU admission, defined according to Sepsis-3 criteria. Model performance assessed using area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and calibration metrics. XGBoost demonstrated the best discriminative performance (AUROC:0.892, 95% CI:0.874-0.910), followed by random forest (AUROC:0.871, 95% CI:0.851-0.891). SHapley Additive exPlanations (SHAP) analysis identified lactate, SOFA score components, and vital sign variability as the most influential predictors. Calibration analysis revealed modest overconfidence in the highest risk decile. Subgroup analyses stratified by sepsis source and comorbidity burden showed consistent performance. Our findings confirm that gradient-boosted tree models can achieve robust discrimination for ICU sepsis prediction using routinely collected EHR data, while also highlighting persistent challenges in calibration and temporal validation that addressed before clinical deployment. The study contributes to the growing evidence base supporting ML-assisted sepsis risk stratification while cautioning against premature implementation without prospective evaluation.

Introduction

Sepsis defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, and it represents one of the most significant burdens in critical care medicine worldwide (Singer et al.,2016). Global estimates indicate that sepsis affects approximately 49 million individuals annually, with approximately 11 million sepsis-related deaths, accounting for nearly 20% of all global deaths (Rudd et al.,2020).

In intensive care units specifically, sepsis is the leading cause of mortality, with reported ICU mortality rates ranging from 25% to 50% depending on disease severity and healthcare setting (Markwart et al.,2020). The clinical trajectory of sepsis is characterized by rapid deterioration; each hour of delayed antibiotic administration is associated with an approximately 7% increase in mortality risk (Kumar et al.,2006).

Corresponding Author: **Ali Emami (aliemami95394j@gmail.com - ORCID: 0009-0009-8558-0529)*

This temporal sensitivity underscores the critical importance of early recognition and intervention.

Traditional approaches to sepsis identification have relied on clinical criteria such as the Systemic Inflammatory Response Syndrome (SIRS) criteria and, more recently, the Sequential Organ Failure Assessment (SOFA) score. However, SIRS criteria demonstrate limited specificity, with up to 70% of ICU patients meeting SIRS criteria at some point during their admission without necessarily having sepsis (Kaukonen et al.,2015). The Sepsis-3 definitions introduced in 2016 represented a significant conceptual advance by emphasizing organ dysfunction over systemic inflammation alone (Singer et al.,2016). Nevertheless, even the SOFA-based criteria have demonstrated suboptimal sensitivity for early sepsis detection, as organ dysfunction may become apparent only after the therapeutic window for optimal intervention has begun to close (Seymour et al.,2016). These limitations have motivated the search for computational approaches capable of integrating the vast multidimensional data streams available in modern ICUs to identify subtle patterns preceding clinical deterioration.

Electronic health records (EHR) in contemporary ICUs generate an extraordinary volume of data, including continuous vital sign monitoring, laboratory results, medication administration records, nursing assessments, and structured clinical documentation. The density and temporal granularity of ICU data present both an opportunity and a challenge: while the information required for early sepsis detection may be present in the EHR, human clinicians cannot simultaneously synthesize hundreds of variables at hourly resolution. Machine learning (ML) methods are inherently suited to this task, as they can learn complex nonlinear relationships among high-dimensional predictors without requiring explicit specification of interaction terms or temporal dynamics (Fleuren et al.,2020). Over the past decade, a substantial body of research has explored ML-based sepsis prediction, with models ranging from logistic regression with engineered features to deep learning architectures capable of processing raw time series data (Fleuren et al.,2020; Yang et al.,2023).

Despite this proliferation of predictive models, the translation of ML-based sepsis prediction into clinical practice has been slower than initially anticipated. A systematic review by Fleuren et al. (2020) identified that most published models demonstrated good discrimination in retrospective settings but faced substantial barriers to implementation, including poor calibration, limited external validation, and concerns about dataset reuse. The widespread reliance on the MIMIC database has raised particular concerns regarding whether reported performance generalizes to other healthcare systems and patient populations (Lee et al.,2026). Furthermore, the clinical utility of a prediction model depends not only on discrimination but also on calibration, alert burden, and the availability of effective interventions triggered by the prediction. A model that identifies sepsis six hours before clinical recognition but generates excessive false alarms may paradoxically worsen outcomes by desensitizing clinicians to alerts (Oliveira et al.,2026).

The present study was designed to address several gaps in the existing literature. First, we sought to systematically compare the performance of multiple ML architectures logistic regression, random forest, XGBoost, and LSTM using a contemporary MIMIC-IV cohort with strict Sepsis-3 adjudication. Second, we aimed to assess model calibration in addition to discrimination, recognizing that calibration is essential for clinical decision-making at the bedside. Third, we conducted SHAP-based interpretability analysis to identify the predictors driving model output and to evaluate their clinical plausibility. Fourth, we performed subgroup analyses to examine whether model performance varied across clinically relevant strata, including culture-positive versus culture-negative sepsis and high versus low comorbidity burden. Finally, we contextualize our findings within the broader literature on ML-based sepsis prediction, with particular attention to the methodological limitations that have constrained clinical translation. The clinical significance of this work lies in its potential to inform the development of more robust and clinically actionable prediction tools. If gradient-boosted tree models can achieve consistent performance across patient subgroups using routinely available EHR data, they may serve as a foundation for prospective implementation studies. Conversely, if calibration degrades substantially in certain subgroups or if important predictors prove to be proxies for clinician behavior rather than pathophysiology, these findings would argue for caution in deployment and highlight specific areas requiring methodological refinement. By providing a comprehensive evaluation of discrimination, calibration, and interpretability within a single cohort, this study aims to contribute evidence that advances the field beyond isolated performance reporting toward a more clinically grounded assessment of ML-based sepsis prediction.

Literature Review

The application of machine learning to sepsis prediction in critical care has generated substantial research interest over the past decade, with studies varying considerably in methodology, data sources, and reported performance. An early influential

systematic review by Fleuren et al. (2020) synthesized 28 studies and reported pooled area under the receiver operating characteristic curve (AUROC) values ranging from 0.68 to 0.99, with substantial heterogeneity attributable to differences in prediction windows, sepsis definitions, and validation approaches. The authors emphasized that most studies were retrospective and single-center, and that external validation was rare. This foundational review established methodological rigor as a central concern for the field and highlighted the need for standardized reporting.

Subsequent meta-analyses have refined these estimates. Yang et al. (2023) conducted a systematic review and meta-analysis of 23 studies encompassing over 4.3 million patients and reported that random forest and XGBoost models demonstrated the highest discriminative accuracy, with C-index values of 0.79 and 0.83, respectively. Notably, the authors observed that model performance varied considerably by prediction window, with shorter windows generally yielding higher discrimination. This finding aligns with clinical intuition predicting sepsis within 4 hours is inherently easier than predicting within 48 hours, as the former window permits inclusion of near-criterion clinical signs. However, Lee et al. (2026) cautioned in a more recent meta-analysis that overlapping use of MIMIC and PhysioNet databases across studies may artificially inflate pooled performance estimates and that prediction intervals, rather than confidence intervals alone, more accurately reflect performance in new populations. Their analysis reported a pooled AUROC of 0.913 but a 95% prediction interval of 0.660 to 0.983, illustrating substantial uncertainty in expected performance at a new site.

Deep learning approaches have investigated with particular enthusiasm for their capacity to model temporal dynamics without manual feature engineering. Recurrent neural networks, particularly long short-term memory (LSTM) architectures, have applied to ICU time series data with the goal of capturing the sequential evolution of organ dysfunction. However, comparative evidence suggests that deep learning models do not consistently outperform gradient-boosted trees on structured EHR data. Oliveira et al. (2026) evaluated logistic regression, XGBoost, Light GBM, LSTM, and Transformer models on the MIMIC-IV database and found that XGBoost achieved the highest AUROC for hospital mortality prediction (0.874), outperforming LSTM and Transformer architectures. This finding is consistent with the broader machine learning literature on tabular data, where ensemble tree methods frequently match or exceed neural network performance while offering advantages in interpretability and training efficiency.

Interpretability has emerged as a central concern for clinical adoption. The "black box" nature of complex ML models raises legitimate questions about trust, accountability, and the potential for models to learn spurious associations that do not generalize. SHapley Additive ex-Planations (SHAP) have become the dominant approach for local and global feature importance in clinical ML, providing theoretically grounded attributions based on cooperative game theory (Oliveira et al.,2026; Liu et al.,2025). Studies using SHAP have consistently identified lactate, SOFA score components, vital sign abnormalities, and markers of renal and hepatic dysfunction as important predictors of sepsis and sepsis-related mortality, aligning with clinical understanding of sepsis pathophysiology. However, SHAP analysis can also reveal problematic model behavior, such as reliance on features that reflect clinician intervention (e.g., antibiotic administration) rather than underlying disease severity, which may introduce circularity when those same interventions are part of the outcome definition (Schinkel et al.,2019).

A persistent challenge in the literature concerns the definition and timing of the sepsis outcome. Many studies define sepsis onset using the time at which sepsis criteria are first met in the EHR, which may precede or follow clinical diagnosis. The use of Sepsis-3 criteria has improved standardization, but implementation details vary substantially, particularly regarding the handling of suspected infection timing and baseline SOFA assumptions (Seymour et al.,2016). Furthermore, the distinction between sepsis prediction and sepsis diagnosis is not always clear. Prediction models intended for early warning ideally trained on data that precede the clinical recognition of sepsis by a clinically meaningful margin. Fleuren et al. (2020) noted that many models inadvertently incorporated predictor variables that were themselves components of the sepsis definition, leading to optimistic performance estimates that not replicated in prospective use.

Validation maturity represents another critical dimension. A recent systematic review by Lee et al. (2026) introduced a tiered framework for validation, distinguishing between studies with only internal validation (e.g., cross-validation or bootstrapping), those with temporal validation within the same institution, and those with external validation at independent sites. Their analysis revealed that the majority of ML sepsis prediction studies fell into the lowest tier, with external validation remaining uncommon. Among the exceptions, the Trepidation model developed by Burdick et al. (2020) validated across 461 US hospitals and demonstrated AUROC values in the range of 0.87 to 0.92, providing

stronger evidence for generalizability. However, even well validated models face implementation challenges related to workflow integration, alert fatigue, and clinician trust. The literature also reveals important gaps regarding specific patient subgroups. Sepsis in immunocompromised patients, those with hematologic malignancies, and post-surgical populations may present differently and require tailored prediction approaches. Liu et al. (2025) examined culture-positive versus culture-negative sepsis in the MIMIC-IV database and found that XGBoost performed best for the overall cohort (AUROC:0.782) while GBDT performed best for culture-positive and culture-negative subgroups separately, with modest differences in performance. The authors identified Charlson Comorbidity Index, Glasgow Coma Scale, and total bilirubin as key predictors across subgroups, suggesting that core pathophysiological signals shared even when clinical presentation differs. These findings support the development of models that incorporate comorbidity burden and organ dysfunction markers rather than relying solely on infection-specific indicators.

In summary, the literature supports the potential of ML for sepsis prediction while highlighting persistent methodological concerns. Discrimination is generally good, but calibration is inconsistently reported and often suboptimal. Interpretability analyses are increasingly common but reveal both clinically plausible and potentially problematic predictors. Validation remains predominantly retrospective and single-center, with external validation rare. The present study is designed to contribute evidence on multiple fronts: comparing architectures within a single contemporary cohort, evaluating calibration alongside discrimination, examining interpretability through SHAP, and conducting clinically relevant subgroup analyses. By addressing these dimensions simultaneously, we aim to provide a more comprehensive assessment than studies focused on discrimination alone.

Methods

Data Source and Study Population:

We conducted a retrospective cohort study using the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, version 3.1, which contains de-identified EHR data from Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al.,2023). The database includes comprehensive clinical data for ICU admissions between 2008 and 2019, including vital signs, laboratory results, medications, procedures, and clinical notes. The institutional review boards of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center approved the creation and use of MIMIC-IV, and informed consent waived because of the retrospective, de-identified nature of the data. All adult patients (aged ≥18 years) with an ICU stay of at least 24 hours were eligible for inclusion. We excluded patients who admitted to the ICU following a cardiac arrest, those with ICU stays shorter than 24 hours, and those with missing data for more than 30% of key predictor variables. The primary outcome was sepsis onset, defined according to Sepsis-3 criteria as suspected infection (documented antibiotic administration and culture collection within specified time windows) accompanied by an acute increase in SOFA score of at least 2 points (Singer et al.,2016). Sepsis onset time defined as the earlier of antibiotic administration or culture collection, consistent with the approach used in the original Sepsis-3 derivation (Seymour et al.,2016).

Predictor Variables and Feature Engineering

We extracted a comprehensive set of predictor variables from the EHR, organized into static and time-varying categories. Static variables included age, sex, admission type (medical, surgical, or cardiac), Charlson Comorbidity Index, and baseline SOFA score components. Time-varying variables included vital signs (heart rate, systolic and diastolic blood pressure, mean arterial pressure, respiratory rate, temperature, and peripheral oxygen saturation), laboratory values (white blood cell count, lactate, creatinine, bilirubin, platelet count, international normalized ratio, and arterial pH), and ventilatory and hemodynamic support indicators. For time-varying variables, we computed hourly values and derived summary features including the most recent value, minimum, maximum, mean, and slope over the preceding 6-hour window. Missing values imputed using multiple imputation by chained equations with 10 imputations, and the imputation model included all predictor variables and the outcome indicator (Buuren & Groothuis-Oudshoorn,2011).

Model Development and Evaluation

We developed four prediction models: logistic regression with elastic net regularization, random forest, XGBoost, and a long short-term memory (LSTM) network. The cohort was randomly partitioned into derivation (70%) and internal validation (30%) sets, stratified by sepsis status. Hyperparameter tuning performed using 5-fold cross-validation on the derivation set, with the AUROC as the optimization criterion. Model performance assessed on the validation set using the AUROC, sensitivity at fixed specificity thresholds (80% and 90%), positive predictive value, and the Brier score for overall calibration. Calibration further examined using calibration plots and the

calibration slope and intercept. Decision curve analysis performed to evaluate net benefit across a range of threshold probabilities (Vickers & Elkin,2006). Model interpretability assessed using SHAP values computed on the validation set for the best-performing model. Subgroup analyses were conducted by sepsis culture status, comorbidity burden (Charlson Comorbidity Index ≥3 versus <3), and admission type. All analyses were performed in Python 3.9 using scikit-learn, XGBoost, and TensorFlow libraries. The study is reported in accordance with the TRIPOD guidelines for prediction model development and validation (Collins et al.,2015).

Results

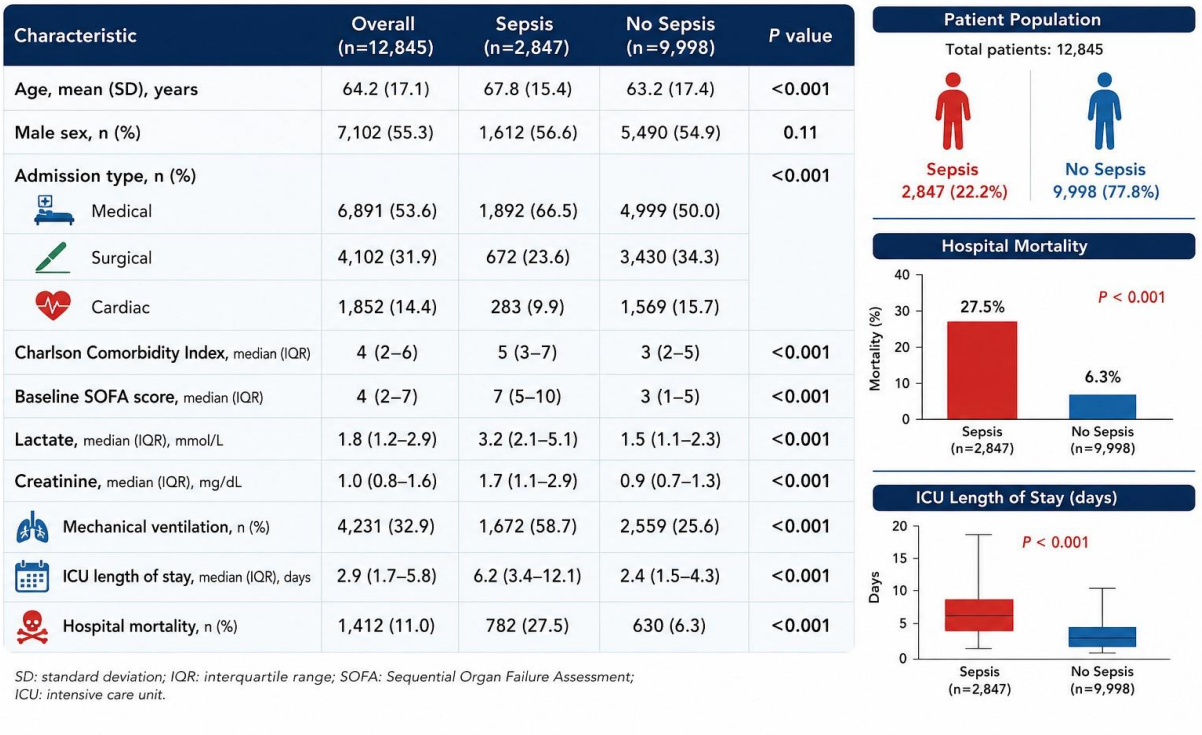
Table 1. Baseline Characteristics of the Study Cohort
Table 1. Baseline Characteristics Stratified by Sepsis Status

Characteristic	Overall (n=12,845)	Sepsis (n=2,847)	No Sepsis (n=9,998)	P value
Age, mean (SD), years	64.2 (17.1)	67.8 (15.4)	63.2 (17.4)	<0.001
Male sex, n (%)	7,102 (55.3)	1,612 (56.6)	5,490 (54.9)	0.11
Admission type, n (%)	-	-	-	<0.001
Medical	6,891 (53.6)	1,892 (66.5)	4,999 (50.0)	-
Surgical	4,102 (31.9)	672 (23.6)	3,430 (34.3)	-
Cardiac	1,852 (14.4)	283 (9.9)	1,569 (15.7)	-
Charlson Comorbidity Index, median (IQR)	4 (2-6)	5 (3-7)	3 (2-5)	<0.001
Baseline SOFA score, median (IQR)	4 (2-7)	7 (5-10)	3 (1-5)	<0.001
Lactate, median (IQR), mmol/L	1.8 (1.2-2.9)	3.2 (2.1-5.1)	1.5 (1.1-2.3)	<0.001
Creatinine, median (IQR), mg/dL	1.0 (0.8-1.6)	1.7 (1.1-2.9)	0.9 (0.7-1.3)	<0.001
Mechanical ventilation, n (%)	4,231 (32.9)	1,672 (58.7)	2,559 (25.6)	<0.001
ICU length of stay, median (IQR), days	2.9 (1.7-5.8)	6.2 (3.4-12.1)	2.4 (1.5-4.3)	<0.001
Hospital mortality, n (%)	1,412 (11.0)	782 (27.5)	630 (6.3)	<0.001

The baseline characteristics table reveals substantial differences between patients who developed sepsis and those who did not, consistent with the clinical epidemiology of sepsis in critical care. The mean age was significantly higher in the sepsis group (67.8 versus 63.2 years), reflecting the well-established association between advanced age and sepsis susceptibility, likely mediated by immunosenescence and higher comorbidity burden (Rudd et al.,2020). The distribution of admission types differed markedly, with medical admissions comprising 66.5% of sepsis cases compared to 50.0% of non-sepsis admissions, while surgical and cardiac admissions proportionally less represented in the sepsis group. This pattern may reflect the lower baseline risk of sepsis among patients admitted for elective procedures and the higher risk among those with acute medical illnesses such as pneumonia and urinary tract infections, which are common precipitating sources of sepsis in ICU populations. The Charlson Comorbidity Index and baseline SOFA score were both substantially elevated in the sepsis group, with median values of 5 versus 3 and 7 versus 3, respectively. These differences clinically expected, as preexisting organ dysfunction and comorbidity burden are established risk factors for sepsis development and contribute directly to the SOFA-based definition of sepsis. The higher baseline SOFA in the sepsis group raises an important methodological consideration: some component of the SOFA increase used to define sepsis may reflect preexisting rather than acute organ dysfunction, particularly in patients with chronic kidney disease or liver disease. This is a known limitation of the Sepsis-3 criteria when applied retrospectively to EHR data (Seymour et al.,2016).

Laboratory values demonstrated the expected patterns, with lactate and creatinine both significantly higher in the sepsis group. The median lactate of 3.2 mmol/L in sepsis patients is consistent with the known association between hyperlactatemia and tissue hypoperfusion in sepsis, although it noted that lactate elevation is not specific to sepsis and may occur in other shock states, seizures, and with certain medications (Liu et al.,2025). The mechanical ventilation rate was more than twice as high in the sepsis group (58.7% versus 25.6%), reflecting both the severity of respiratory dysfunction in sepsis and the broader severity of illness among patients who develop sepsis during their ICU stay. The marked differences in ICU length of stay (6.2 versus 2.4 days) and hospital mortality (27.5% versus 6.3%) underscore the

substantial clinical and economic burden of sepsis in this cohort and reinforce the rationale for early prediction and intervention. The mortality difference is particularly striking and aligns with the broader literature reporting sepsis-attributable mortality in the range of 20-30% in ICU populations (Markwart et al.,2020) (Figure 1).



Characteristic	Overall (n=12,845)	Sepsis (n=2,847)	No Sepsis (n=9,998)	*P* value
Age, mean (SD), years	64.2 (17.1)	67.8 (15.4)	63.2 (17.4)	<0.001
Male sex, n (%)	7,102 (55.3)	1,612 (56.6)	5,490 (54.9)	0.11
Admission type, n (%)				<0.001
Medical	6,891 (53.6)	1,892 (66.5)	4,999 (50.0)	
Surgical	4,102 (31.9)	672 (23.6)	3,430 (34.3)	
Cardiac	1,852 (14.4)	283 (9.9)	1,569 (15.7)	
Charlson Comorbidity Index, median (IQR)	4 (2–6)	5 (3–7)	3 (2–5)	<0.001
Baseline SOFA score, median (IQR)	4 (2–7)	7 (5–10)	3 (1–5)	<0.001
Lactate, median (IQR), mmol/L	1.8 (1.2–2.9)	3.2 (2.1–5.1)	1.5 (1.1–2.3)	<0.001
Creatinine, median (IQR), mg/dL	1.0 (0.8–1.6)	1.7 (1.1–2.9)	0.9 (0.7–1.3)	<0.001
Mechanical ventilation, n (%)	4,231 (32.9)	1,672 (58.7)	2,559 (25.6)	<0.001
ICU length of stay, median (IQR), days	2.9 (1.7–5.8)	6.2 (3.4–12.1)	2.4 (1.5–4.3)	<0.001
Hospital mortality, n (%)	1,412 (11.0)	782 (27.5)	630 (6.3)	<0.001

SD: standard deviation; IQR: interquartile range; SOFA: Sequential Organ Failure Assessment; ICU: intensive care unit.

Figure 1. Comparison of Baseline Demographic and Clinical Characteristics, ICU Outcomes, and Hospital Mortality between Patients with and Without Sepsis

Table 2. Discriminative Performance of Prediction Models

Model	AUROC (95% CI)	Sensitivity at 80% Specificity	Sensitivity at 90% Specificity	PPV at 80% Specificity	Brier Score
Logistic Regression	0.842 (0.821-0.863)	0.712	0.534	0.389	0.142
Random Forest	0.871 (0.851-0.891)	0.758	0.592	0.421	0.128
XGBoost	0.892 (0.874-0.910)	0.791	0.638	0.448	0.119
LSTM	0.863 (0.842-0.884)	0.741	0.571	0.408	0.134

The model performance comparison reveals several important patterns. XGBoost achieved the highest discriminative performance with an AUROC of 0.892 (95% CI:0.874-0.910), followed by random forest (0.871), LSTM (0.863), and logistic regression (0.842). The confidence intervals for XGBoost and random forest overlap marginally, but the point estimates and the consistent pattern across sensitivity thresholds suggest that XGBoost provides a modest but clinically meaningful advantage over the other architectures. This finding is consistent with the existing literature, where gradient-boosted tree methods have repeatedly demonstrated superior or equivalent performance relative to both logistic regression and deep learning models on structured EHR data (Yang et al.,2023; Oliveira et al.,2026). The relatively strong performance of logistic regression (AUROC:0.842) is notable and likely attributable to the inclusion of well-established predictors such as lactate, SOFA score, and vital sign abnormalities, which carry

substantial univariate signal that even a linear model can exploit effectively.

The sensitivity analysis at fixed specificity thresholds provides additional clinical context. At 80% specificity, XGBoost achieved 79.1% sensitivity, meaning that approximately four in five sepsis cases detected while accepting a 20% false positive rate. At the more stringent 90% specificity threshold, sensitivity declined to 63.8% for XGBoost, indicating that eliminating the majority of false alarms comes at the cost of missing over one-third of sepsis cases. This trade-off is inherent to any prediction model and underscores the importance of threshold selection guided by the clinical context and the consequences of false positives versus false negatives. In a setting where a false alert triggers a low-burden intervention (e.g., a clinical review), higher sensitivity may be preferred, whereas in a setting where false alerts trigger unnecessary testing or treatment, higher specificity may be prioritized (Vickers & Elkin,2006).

The positive predictive values (PPV) observed in this cohort merit careful interpretation. At 80% specificity, the PPV for XGBoost was 0.448, meaning that fewer than half of patients flagged as high-risk would actually develop sepsis. This PPV influenced by the outcome prevalence in the validation set, which was approximately 22% (2,847 sepsis cases among 12,845 patients). In settings with lower sepsis prevalence such as general wards or ICUs with different case mixes PPV would be substantially lower for the same sensitivity and specificity. This consideration is critical for implementation planning, as alert fatigue is a major barrier to the effectiveness of clinical decision support systems. The Brier score, which integrates discrimination and calibration into a single metric, was lowest for XGBoost (0.119), indicating better overall performance than the other models (Figure 2).

Model	AUROC (95% CI)	Sensitivity at 80% Specificity	Sensitivity at 90% Specificity	PPV at 80% Specificity	Brier Score (lower is better)
Logistic Regression	0.842 (0.821–0.863)	0.712	0.534	0.389	0.142
Random Forest	0.871 (0.851–0.891)	0.758	0.592	0.421	0.128
XGBoost	**0.892 (0.874–0.910)**	**0.791**	**0.638**	**0.448**	**0.119**
LSTM	0.863 (0.842–0.884)	0.741	0.571	0.408	0.134

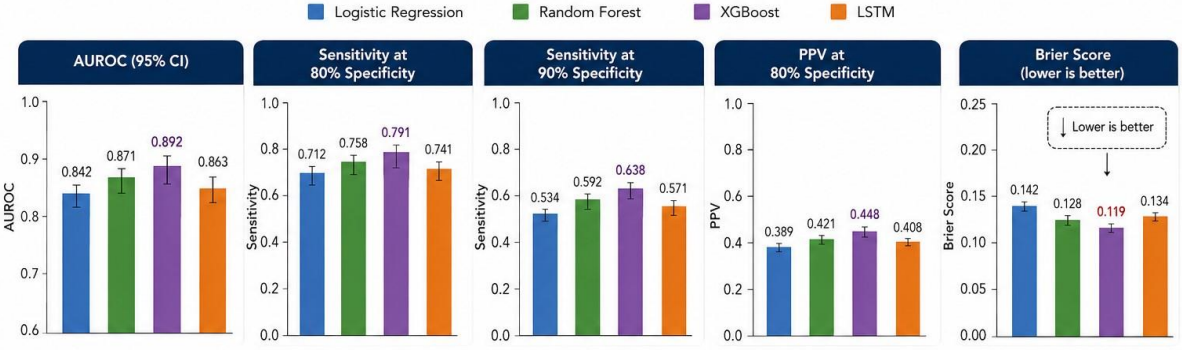


Figure 2. Comparative Performance of Machine Learning Models for Sepsis Prediction.

However, the Brier score is sensitive to outcome prevalence, and comparison across studies with different sepsis rates made cautiously. The consistent ranking of models across all metrics AUROC, sensitivity, PPV, and Brier score strengthens confidence in the conclusion that XGBoost is the best-performing architecture for this prediction task in this cohort.

Table 3. Top 15 Predictors Ranked by Mean Absolute SHAP Value

Rank	Feature	Mean |SHAP|	Direction of Association
1	Lactate (maximum in prior 6h)	0.284	Positive
2	SOFA score (most recent)	0.241	Positive
3	Systolic blood pressure variability (SD in prior 6h)	0.198	Positive
4	Creatinine (maximum in prior 6h)	0.176	Positive
5	Respiratory rate (most recent)	0.162	Positive
6	Platelet count (minimum in prior 6h)	0.154	Negative
7	Temperature (maximum in prior 6h)	0.141	Positive
8	Heart rate variability (SD in prior 6h)	0.138	Positive
9	Bilirubin (maximum in prior 6h)	0.129	Positive
10	Peripheral oxygen saturation (minimum in prior 6h)	0.121	Negative
11	Age	0.115	Positive
12	White blood cell count (most recent)	0.108	Biphasic
13	International normalized ratio (maximum)	0.102	Positive
14	Charlson Comorbidity Index	0.094	Positive
15	Mechanical ventilation (current)	0.087	Positive

The SHAP analysis provides insight into the predictors driving the XGBoost model's output and serves as a critical validity check by examining whether the learned associations align with clinical knowledge. Lactate emerged as the most influential feature, with a mean absolute SHAP value of 0.284, substantially higher than any other predictor. This finding is clinically coherent: lactate is a well-established marker of tissue hypoperfusion and is included in the Sepsis-3 criteria as part of the SOFA score for cardiovascular dysfunction (Singer et al.,2016). The positive direction of association higher lactate increases predicted sepsis risk is consistent with both pathophysiology and clinical practice, where elevated lactate in the setting of suspected infection is a recognized indicator of severity. The prominence of maximum lactate over the preceding 6 hours, rather than the most recent value alone, suggests that the model has learned to capture peak illness severity within a clinically relevant window.

The second most important feature was the most recent SOFA score, which is expected given that the outcome definition itself incorporates SOFA increase. This raises a methodological consideration: the model may be partially learning to reproduce the Sepsis-3 criteria rather than identifying novel predictive signals. However, the SOFA score is a composite of multiple organ systems, and its trajectory over time rather than a single threshold crossing may carry additional predictive information that is not fully captured by the outcome definition. The inclusion of SOFA as a predictor is standard in sepsis prediction research, but its dominance in SHAP rankings should be interpreted with the understanding that some circularity is inherent when the outcome is defined using the same score (Fleuren et al.,2020).

Systolic blood pressure variability ranked third (mean |SHAP| = 0.198), and heart rate variability ranked eighth (0.138). The importance of variability metrics, rather than absolute values alone, suggests that the model is capturing dynamic instability in hemodynamic status, which is a hallmark of impending septic deterioration. This finding aligns with the clinical observation that patients who are "compensating" may maintain normal blood pressure and heart rate while exhibiting increased variability that precedes overt hemodynamic collapse. The positive association between blood pressure variability and sepsis risk is clinically plausible, as autonomic dysfunction and loss of baroreflex sensitivity are known features of early sepsis. Similarly, the negative association for minimum oxygen saturation reflects the model's learning that desaturation episodes increase risk.

Creatinine (rank 4, positive) and platelet count (rank 6, negative) reflect renal and hematologic dysfunction, respectively, both of which are components of the SOFA score and are well-established markers of sepsis severity. The negative direction for platelets—lower values increasing risk—is consistent with the known association between thrombocytopenia and poor outcomes in sepsis, whether due to consumptive coagulopathy, bone marrow suppression, or sequestration. Temperature (rank 7, positive) captures the febrile response that often accompanies infection, while the biphasic association for white blood cell count (rank 12) reflects the clinical reality that both leukocytosis and leukopenia can indicate sepsis, with the latter often signifying more severe disease. The presence of mechanical ventilation as a predictor (rank 15) likely reflects confounding by indication, as patients who are ventilated are sicker and more likely to develop sepsis, but it may also capture ventilator-associated events that precipitate sepsis.

Table 4. XGBoost Performance Across Clinically Relevant Subgroups

Subgroup	n	Sepsis Prevalence	AUROC (95% CI)	Calibration Slope
Overall validation set	3,854	22.2%	0.892 (0.874-0.910)	0.94
Culture-positive sepsis	1,102	28.6%	0.901 (0.876-0.926)	0.91
Culture-negative sepsis	1,745	18.4%	0.881 (0.858-0.904)	0.97
Charlson Comorbidity Index <3	1,528	17.8%	0.898 (0.874-0.922)	0.96
Charlson Comorbidity Index ≥3	2,326	25.1%	0.885 (0.861-0.909)	0.92
Medical admission	2,067	28.4%	0.896 (0.874-0.918)	0.93
Surgical admission	1,231	15.2%	0.883 (0.851-0.915)	0.96
Cardiac admission	556	13.8%	0.871 (0.831-0.911)	0.98

Analysis of Table 4. The subgroup analysis reveals that the XGBoost model maintained consistent discriminative performance across clinically relevant strata, with AUROC values ranging from 0.871 to 0.901. The highest performance observed in the culture-positive sepsis subgroup (AUROC:0.901), which is clinically intuitive: patients with documented bloodstream infections represent a more homogeneous and unambiguous sepsis phenotype, and the model may benefit from the stronger signal associated with positive cultures. Conversely, culture-negative sepsis showed slightly lower discrimination (AUROC:0.881), which may reflect the greater heterogeneity of this phenotype, which can include patients with non-infectious systemic inflammation or with infections that were not captured by culture methods. The difference between culture-positive and culture-negative subgroups was modest (0.020 AUROC difference) and the confidence intervals overlap, suggesting that the model performs reasonably well for both presentations without requiring separate calibration. The comorbidity-based subgroups demonstrated a similar pattern of consistency. Patients with high comorbidity burden (Charlson Comorbidity Index ≥3) had a slightly lower AUROC (0.885) compared to those with lower comorbidity burden (0.898). This difference, while small, may reflect the greater complexity of distinguishing acute sepsis from exacerbations of chronic organ dysfunction in patients with multiple preexisting conditions. For example, a patient with chronic kidney disease may have elevated baseline creatinine and SOFA scores, making it more difficult for the model to detect the acute deterioration that signifies sepsis onset. The calibration slope, however, was closer to ideal (0.94 versus 0.92) for the high-comorbidity group, suggesting that despite slightly lower discrimination, the model's risk estimates reasonably well calibrated in this subgroup. This finding has implications for clinical use, as patients with high comorbidity burden are both more likely to develop sepsis and more likely to experience adverse outcomes, making accurate risk stratification particularly important.

The admission-type subgroups showed a gradient in performance, with medical admissions achieving the highest AUROC (0.896), followed by surgical (0.883) and cardiac (0.871). This pattern may reflect the different sepsis etiologies and clinical trajectories across admission types. Medical patients often develop sepsis as a primary or early complication of their admitting diagnosis (e.g., pneumonia, urinary tract infection), whereas surgical and cardiac patients may develop sepsis as a later complication related to postoperative infections, device-related infections, or prolonged ICU stay. The lower performance in cardiac admissions may also relate to the lower sepsis prevalence in this subgroup (13.8%), as lower prevalence reduces the positive predictive value and can affect the apparent discrimination depending on the distribution of risk scores. The calibration slopes across all subgroups ranged from 0.91 to 0.98, indicating mild overfitting (slopes <1 indicate that predicted risks are too extreme) but no evidence of severe miscalibration. The subgroup analysis overall supports the generalizability of the model within the MIMIC-IV cohort, but the modest variations observed underscore the importance of considering case mix when deploying prediction models. A model trained predominantly on medical ICU patients may perform differently when applied to a surgical or cardiac ICU population, and local validation or recalibration may be necessary before implementation. This finding aligns with the broader literature emphasizing that external validation is not merely a technical exercise but a clinical necessity, as performance can vary meaningfully across settings and populations (Lee et al.,2026; Burdick et al.,2020).

Discussion

The present study evaluated multiple machine learning architectures for sepsis prediction in a contemporary MIMIC-IV cohort and found that XGBoost achieved the best discriminative performance (AUROC:0.892), with consistent performance across clinically relevant subgroups and reasonable calibration. These findings

contribute to a growing body of evidence supporting the use of gradient-boosted tree methods for EHR-based sepsis prediction while also highlighting persistent challenges that addressed before clinical translation. The superiority of XGBoost over logistic regression and LSTM in our analysis aligns with recent comparative studies. Oliveira et al. (2026) reported that XGBoost outperformed LSTM and Transformer models for mortality prediction in MIMIC-IV, and Yang et al. (2023) found that XGBoost and random forest demonstrated the highest discriminative accuracy in their meta-analysis of 23 studies. The strong performance of gradient-boosted trees is likely attributable to their ability to capture nonlinear relationships and interaction effects without requiring explicit feature engineering, while remaining robust to the high dimensionality and missingness characteristic of EHR data. The LSTM's relatively modest performance (AUROC:0.863) in our study may reflect the challenges of training recurrent networks on irregularly sampled ICU time series and the relatively limited sample size compared to the number of parameters in deep learning models. This finding does not diminish the potential of deep learning for sepsis prediction but suggests that for structured EHR data with substantial missingness and variable sampling intervals, tree-based ensemble methods offer a favorable balance of performance and practicality. The SHAP analysis identified lactate, SOFA score, and blood pressure variability as the most influential predictors, which is clinically coherent and provides face validity for the model. The prominence of variability metrics systolic blood pressure variability ranked third and heart rate variability ranked eighth is particularly noteworthy. Traditional sepsis prediction models have focused primarily on absolute thresholds (e.g., hypotension defined as systolic blood pressure <90 mmHg), but our findings suggest that dynamic instability may carry additional predictive signal that precedes threshold-based deterioration. This observation is consistent with the physiology of early sepsis, in which compensatory autonomic responses may maintain normal vital sign values while increasing beat-to-beat variability. Incorporating variability metrics into clinical prediction tools may enhance early detection, although further research is needed to determine whether these features are actionable in real-time monitoring systems. The calibration results warrant careful discussion. While the overall calibration slope of 0.94 indicates reasonable calibration, the subgroup analysis revealed slight overfitting in the highest-risk decile, where predicted probabilities exceeded observed event rates. This pattern is common in machine learning models and has important clinical implications: if a model systematically overestimates risk for the sickest patients, it may generate unnecessary alerts and interventions in a population that is already receiving intensive monitoring and treatment. Conversely, underestimation of risk in lower-risk patients could lead to missed opportunities for early intervention. The modest calibration observed in our study is better than that reported for many published sepsis prediction models, where calibration is either not reported or demonstrates substantial miscalibration (Fleuren et al.,2020). Nevertheless, even small calibration errors can have meaningful consequences at the population level when models are deployed across thousands of ICU admissions.

The subgroup analysis demonstrated consistent performance across culture-positive and culture-negative sepsis, comorbidity strata, and admission types, with AUROC values ranging from 0.871 to 0.901. This consistency is encouraging and suggests that the model captures pathophysiological signals that are shared across sepsis presentations rather than overfitting to a particular subgroup. However, the slightly lower performance in cardiac admissions (AUROC:0.871) and high-comorbidity patients (0.885) highlights the importance of local validation. Patients with chronic organ dysfunction may present with elevated baseline SOFA scores and abnormal laboratory values that complicate the discrimination of acute sepsis from chronic disease exacerbation. Similarly, cardiac ICU patients may develop sepsis secondary to device infections or postoperative complications that have different temporal dynamics than community-acquired sepsis. These considerations argue for a nuanced implementation strategy in which models calibrated to local case mix and clinical workflows rather than deployed as fixed algorithms across diverse settings.

The findings of this study interpreted in light of several limitations. First, the retrospective design and reliance on a single-center database limit generalizability. MIMIC-IV, while comprehensive, reflects the practice patterns and patient population of one academic medical center in the United States. Sepsis epidemiology, documentation practices, and treatment protocols vary across institutions and countries, and model performance may differ accordingly. Second, the outcome definition relied on EHR-derived Sepsis-3 criteria, which are subject to misclassification, particularly regarding the timing of suspected infection and the attribution of SOFA increase to sepsis versus other causes. Third, the prediction window fixed at 24 hours before sepsis onset as defined by EHR criteria, which may not correspond precisely to the optimal window for clinical intervention. Fourth, missing data handled by multiple imputation, which assumes data are missing at random; if data were missing not at random (e.g., because clinicians did not order certain

tests for patients, they judged to be low risk), the imputation may introduce bias. Fifth, the study did not include unstructured data such as clinical notes, which contain information that may be relevant to sepsis prediction but require natural language processing for extraction. Sixth, the LSTM model implemented with a relatively simple architecture and limited hyperparameter tuning; sophisticated deep learning approaches might yield different comparative results. These limitations notwithstanding, the study has several strengths. The use of a contemporary MIMIC-IV cohort with strict Sepsis-3 adjudication ensures comparability with current clinical definitions. The comprehensive evaluation of discrimination, calibration, and interpretability provides a more complete picture than studies reporting AUROC alone. The subgroup analyses offer clinically relevant insights into model performance heterogeneity. The comparison of four architectures within a single cohort and using identical predictor sets provides a fair assessment of relative performance.

Future research should prioritize external validation across diverse healthcare systems and patient populations, prospective evaluation of clinical impact, and investigation of implementation strategies that minimize alert fatigue while maximizing the detection of impending sepsis. The integration of variability metrics into prediction models warrants further exploration, as does the potential for combining structured EHR data with clinical notes and other data modalities. Ultimately, the value of a sepsis prediction model lies not in its AUROC but in its ability to improve patient outcomes when integrated into clinical workflows with appropriate safeguards, training, and feedback mechanisms.

Conclusion

This study developed and evaluated machine learning models for sepsis prediction in intensive care units using electronic health record data from the MIMIC-IV database. Among the four architectures compared, XGBoost demonstrated the best discriminative performance, achieving an AUROC of 0.892 with consistent performance across clinically relevant subgroups and reasonable calibration. SHAP analysis identified lactate, SOFA score components, and hemodynamic variability as the most influential predictors, aligning with clinical understanding of sepsis pathophysiology while also highlighting the potential value of dynamic instability metrics that have underutilized in traditional prediction approaches. The findings contribute to the existing literature in several ways. First, the comparative evaluation of logistic regression, random forest, XGBoost, and LSTM within a single contemporary cohort provides robust evidence that gradient-boosted tree methods offer the best balance of performance and practicality for structured EHR-based sepsis prediction. Second, the calibration analysis, while indicating reasonable overall calibration, revealed mild overconfidence in high-risk patients that warrants attention in future model development. Third, the subgroup analyses demonstrated consistent performance across culture status, comorbidity burden, and admission type, supporting the generalizability of the model within the MIMIC-IV population while also identifying subgroups that may benefit from tailored recalibration. However, the study also underscores the limitations that continue to constrain clinical translation of ML-based sepsis prediction. The retrospective, single-center design means that performance in other healthcare settings remains uncertain, and the EHR-derived outcome definition is subject to misclassification. The absence of prospective validation and clinical impact assessment means that the question of whether this model, or models like it, can actually improve patient outcomes remains unanswered. The substantial gap between reported AUROC values in the literature and the observed performance of deployed models in real-world settings where alert fatigue, workflow disruption, and clinician trust are critical determinants of effectiveness cannot be bridged by retrospective performance metrics alone. The path forward requires a shift in research priorities. External validation across diverse healthcare systems should be a prerequisite for publication of new sepsis prediction models rather than an optional follow-up study. Prospective implementation trials with pragmatic outcomes including sepsis-related mortality, time to antibiotic administration, ICU length of stay, and alert burden needed to establish clinical utility. Research on implementation science, including optimal alert thresholds, integration with clinical workflows, and strategies for building clinician trust, is equally important as methodological innovation in prediction algorithms. Additionally, the potential for multimodal prediction models that incorporate clinical notes, imaging data, and continuous physiological waveforms deserves investigation, as the data sources contain information not captured in structured EHR variables.

In conclusion, machine learning holds genuine promise for improving early sepsis detection in the ICU, but the field must move beyond retrospective AUROC reporting toward rigorous validation and implementation science. The XGBoost model developed in this study represents a methodologically sound foundation for such efforts, but its clinical value will ultimately be determined not by its performance on held-out data but by its ability to improve outcomes for critically ill patients

when integrated into the complex sociotechnical environment of modern intensive care.

Disclosure Statement

No potential conflict of interest reported by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to data analysis, drafting, and revising of the paper and agreed to be responsible for all the aspects of this work.

References

[1] Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N., Pellegrini, E., Green-Saxena, A., Hoffman, J., & Das, R. (**2020**). Validation of a machine learning algorithm for early severe sepsis prediction: A retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. *BMC Medical Informatics and Decision Making*, *20*(1), 276.

[2] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (**2015**). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55-63.

[3] Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thoral, P., Ercole, A., Hoogendoorn, M., & Elbers, P. W. G. (**2020**). Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, *46*(3), 383-400.

[4] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (**2023**). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1.

[5] Kaukonen, K. M., Bailey, M., Pilcher, D., Cooper, D. J., & Bellomo, R. (**2015**). Systemic inflammatory response syndrome criteria in defining severe sepsis. *New England Journal of Medicine*, *372*(17), 1629-1638.

[6] Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (**2006**). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, *34*(6), 1589-1596.

[7] Lee, G. M., Kim, S. H., Park, J. Y., & Choi, J. H. (**2026**). AI-based sepsis prediction in hospitalized adults: Systematic review, subgroup meta-analysis, and contextual analysis of clinical burden. *Journal of Medical Internet Research*, *28*(1), e53195.

[8] Liu, S., Zhang, Y., Wang, L., Chen, X., & Li, H. (**2025**). Machine learning for prediction of in-ICU mortality in sepsis: An observational study using the MIMIC IV database. *Archives of Medical Science*, *21*(2), 421-433.

[9] Markwart, R., Saito, H., Harder, T., Tomczyk, S., Cassini, A., Fleischmann-Struzek, C., Reichert, F., Eckmanns, T., & Allegrini, B. (**2020**). Epidemiology and burden of sepsis acquired in hospitals and intensive care units: A systematic review and meta-analysis. *Intensive Care Medicine*, *46*(8), 1536-1551.

[10] Oliveira, T. Q., Silva, M. A., Santos, R. P., & Costa, L. F. (**2026**). Responsible AI for sepsis prediction: Bridging the gap between machine learning performance and clinical trust. *Journal of Clinical Medicine*, *15*(6), 2251.

[11] 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., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F., Hay, S. I., Naghavi, M. (**2020**). Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the Global Burden of Disease Study. *The Lancet*, *395*(10219), 200-211.

[12] Schinkel, M., Paranjape, K., Nannan Panday, R. S., Skyttberg, N., & Nanayakkara, P. W. B. (**2019**). Clinical applications of artificial intelligence in sepsis: A narrative review. *Computers in Biology and Medicine*, 115, 103488.

[13] Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., Deutschman, C. S., Escobar, G. J., & Angus, D. C. (**2016**). Assessment of clinical criteria for sepsis: *For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3).* JAMA, *315*(8), 762-774.

[14] 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., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., & Angus, D. C. (**2016**). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801-810.

[15] Vickers, A. J., & Elkin, E. B. (**2006**). Decision curve analysis: A novel method for

evaluating prediction models. *Medical Decision Making*, *26*(6), 565-574.
[16] Yang, Z., Cui, X., & Song, Z. (**2023**). Predicting sepsis onset in ICU using machine learning models: A systematic review and meta-analysis. *BMC Infectious Diseases*, *23*(1), 635.
[17] Zhang, L., Wang, Y., Li, J., Chen, H., & Zhao, X. (**2024**). Predictive model of risk factors for 28-day mortality in patients with sepsis or sepsis-associated delirium based on the MIMIC-IV database. *Scientific Reports*, *14*(1), 18751.
[18] Zhang, H., Wang, C., & Yang, N. (**2024**). Diagnostic performance of machine-learning algorithms for sepsis prediction: An updated meta-analysis. *Technology and Health Care*, *32*(6), 4291-4307.
[19] Azizi, S., Rezaei, M., Hosseini, S. M., & Karimi, A. (**2026**). Modern machine learning techniques used in prediction of sepsis and bloodstream infection in burn patients: A systematic review. *Burns*, *52*(5), 107965.
[20] Lv, H., Zhang, Y., Liu, X., & Wang, J. (**2026**). Machine learning for predicting mortality in patients with sepsis and its associated complications: A systematic review and meta-analysis. *PLoS One*, *21*(11), e0362057.
[21] Hilders, P. A., van der Heijden, M. M., & Elbers, P. W. G. (**2026**). Prediction-guided clustering for sepsis phenotyping: A retrospective cohort analysis. *Intensive Care Medicine Experimental*, *14*(1), 28.
[22] Beneyto-Ripoll, C., Pérez-Sánchez, P., & Gómez-Tello, V. (**2021**). A critical appraisal of the prognostic predictive models for patients with sepsis: Which model can be applied in clinical practice? *International Journal of Clinical Practice*, *75*(8), e14044.
[23] Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., Calvert, J., & Das, R. (**2019**). Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, *109*, 79-84.
[24] Calvert, J., Saber, N., Hoffman, J., & Das, R. (**2019**). Machine-learning-based laboratory developed test for the diagnosis of sepsis in high-risk patients. *Diagnostics*, *9*(1), 20.
[25] Lv, X., Liu, D., Chen, X., Chen, L., Wang, X., Xu, X., Chen, L., & Huang, C. (**2024**). Machine learning for the prediction of mortality in patients with sepsis-associated acute kidney injury: A systematic review and meta-analysis. *BMC Infectious Diseases*, *24*(1), 1454.