RESEARCH PAPER

Prediction of In-Hospital Mortality in Sepsis Using Advanced Predictive Models: Comparison with Traditional Severity Scores

[1]Karan Jain, [2]P. B. Bharate, [3]Amol Patil, [4]Satish Wadde, [5]Amol Kinge and [6]Dilesh Bagul

[1]Research Scholar, Department of Statistics, Malwanchal University, Indore, MP, India
[2]Professor, Department of Statistics, Malwanchal University, Indore, MP, India
[3]Assistant Professor- Statistician, Department of Community Medicine, Government Medical College, Nashik, Maharashtra, India
[4]Associate Professor, JBMGMC Nandurbar, Maharashtra, India
[5]Assistant Professor, SBHGMC Dhule, Maharashtra, India
[6]Officer Biomedical Statistician, Tata Main Hospital, Jamshedpur, Jharkhand, India
***Corresponding Author:** Karan Jain, Research Scholar, Department of Community Medicine, Malwanchal University, Indore, MP, India*
magicofkaran@gmail.com

Received: 16th Dec, 2025; Revised: 8th Feb 2026; Accepted: 12th Feb, 2026; Available Online: 28th Feb, 2026

ABSTRACT

Background: Sepsis remains a leading cause of in-hospital mortality worldwide, and early identification of high-risk patients is essential for timely intervention. Traditional severity scoring systems such as SOFA and APACHE II have limited predictive accuracy in heterogeneous clinical settings. Machine-learning (ML) approaches offer the potential to improve mortality prediction by modeling complex, nonlinear relationships in routinely collected clinical data.

Methods: This retrospective observational study included 500 adult patients admitted with sepsis to a tertiary care hospital. In-hospital mortality was the primary outcome. Demographic characteristics, clinical variables, laboratory parameters, comorbidities, treatment-related factors, and established severity scores were extracted from medical records. Multiple ML models, including logistic regression, naïve Bayes, k-nearest neighbors, support vector machine, random forest, AdaBoost, and extreme gradient boosting (XGBoost), were developed and evaluated. Model performance was assessed using the area under the receiver operating characteristic curve (AUROC), calibration measures, and decision curve analysis, and was compared with traditional severity scores.

Results: Of the 500 patients, 170 (34%) died while 330 (66%) survived during hospitalization. Non-survivors were older and had higher severity scores and worse physiological and laboratory parameters at admission. Among the ML models, XGBoost demonstrated the best performance, achieving the highest AUROC and superior calibration compared with other ML algorithms and traditional scores. The XGBoost model outperformed SOFA, APACHE II, and NEWS2 in predicting in-hospital mortality. Feature importance analysis identified serum lactate, SOFA score, renal dysfunction, hypotension, vasopressor requirement, thrombocytopenia, and age as key predictors of mortality.

Conclusion: Machine-learning models, particularly gradient-boosting approaches, provide more accurate prediction of in-hospital mortality in sepsis than conventional severity scores. These findings support the potential role of ML-based tools in early risk stratification and personalized decision support for sepsis care.

Keywords: Sepsis; In-hospital mortality; Machine learning; XGBoost; Severity scoring systems; Mortality prediction; Clinical decision support

How to cite this article: Jain K, Bharate PB, Patil A, Wadde S, Kinge A, Bagul D, Prediction of In-Hospital Mortality in Sepsis Using Advanced Predictive Models: Comparison with Traditional Severity Scores. Int J Drug Deliv Technol. 2026; 16(2): 435-441; DOI: 10.25258/ijddt.16.2.48

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Sepsis is a life-threatening clinical syndrome arising from a dysregulated host response to infection and remains a major cause of in-hospital mortality worldwide. Despite advances in antimicrobial therapy, organ support, and intensive care practices, sepsis continues to account for a substantial proportion of global deaths. The Global Burden of Disease study estimated nearly 49 million sepsis cases and 11 million sepsis-related deaths worldwide in 2017, representing approximately 20% of all global mortality [1]. These findings highlight the persistent challenge of early identification of high-risk patients and the need for accurate prognostic tools to guide clinical decision-making.

**Author for Correspondence: magicofkaran@gmail.com*

The modern conceptual framework of sepsis was redefined by the Sepsis-3 consensus, which emphasized organ dysfunction rather than systemic inflammation as the defining feature of sepsis [2]. The Sequential Organ Failure Assessment (SOFA) score was proposed as a standardized method for quantifying organ dysfunction and predicting mortality. Alongside SOFA, other traditional severity scoring systems such as the Acute Physiology and Chronic Health Evaluation II (APACHE II) score have been widely used for mortality prediction in critically ill patients [3]. While these scores provide valuable population-level prognostic information, their predictive performance is often limited by static design, linear assumptions, and reduced accuracy when applied to heterogeneous patient populations or early stages of disease [4].

The widespread adoption of electronic health records has enabled access to large volumes of routinely collected clinical data, facilitating retrospective analyses and the development of data-driven prognostic models. The creation of publicly available critical-care databases such as MIMIC-III marked a major milestone in reproducible critical-care research and has enabled the development and validation of advanced prediction models for sepsis outcomes [5]. Retrospective clinical datasets capture complex interactions between demographics, vital signs, laboratory parameters, and treatment interventions, providing a rich substrate for advanced analytical techniques.

Machine learning (ML) methods offer a flexible framework for modeling nonlinear relationships and high-dimensional interactions that are difficult to capture using conventional statistical approaches. Several original cohort studies have demonstrated superior discrimination of ML-based models compared with traditional severity scores for predicting mortality in sepsis. Random forest and gradient-boosting algorithms, in particular, have shown improved performance by integrating heterogeneous clinical variables without prespecified assumptions [6,7]. These findings suggest that ML models may provide more accurate individualized risk estimates in complex clinical conditions such as sepsis.

More recent studies have explored deep-learning approaches for early detection and outcome prediction in sepsis using longitudinal clinical data. Recurrent neural network–based models have demonstrated the ability to detect sepsis and predict mortality earlier than conventional clinical criteria, potentially allowing timely intervention [8]. Additionally, the incorporation of unstructured clinical notes alongside structured data has further improved predictive accuracy, underscoring the versatility of ML techniques in extracting clinically meaningful patterns from electronic health records [9].

Despite these advances, several challenges hinder the translation of ML models into routine clinical practice. Concerns regarding overfitting, calibration, generalizability across institutions, and lack of interpretability remain significant barriers. Comparative evaluations that benchmark ML algorithms against established severity scores using robust performance metrics—including discrimination, calibration, and clinical utility—are essential to establish their real-world value [10]. Moreover, explainable ML approaches are increasingly emphasized to enhance clinician trust and facilitate adoption in high-stakes environments such as critical care.

In this context, the present study aims to develop and evaluate multiple machine-learning models for predicting in-hospital mortality among patients admitted with sepsis and to compare their performance with traditional severity scoring systems. By leveraging retrospective clinical data and employing comprehensive model evaluation and interpretability techniques, this study seeks to assess whether ML-based approaches can offer reliable and clinically meaningful improvements over conventional prognostic tools.

METHODOLOGY

Study Design and Setting

This study was a retrospective observational analysis conducted using routinely collected clinical data from a tertiary care hospital. The study design focused on developing and evaluating predictive models for in-hospital mortality among patients diagnosed with sepsis using machine learning techniques and comparing their performance with traditional severity scoring systems.

Study Population

All adult patients admitted to the hospital with a diagnosis of sepsis during the study period were screened for inclusion. Sepsis was defined according to the Sepsis-3 criteria, based on suspected or documented infection with evidence of organ dysfunction. Patients aged 18 years or older with complete outcome data for in-hospital mortality were included. Patients with incomplete admission records, duplicate entries, or missing outcome status were excluded from the analysis.

Outcome Definition

The primary outcome of interest was in-hospital mortality, defined as death occurring at any time during the index hospital admission. Patients discharged alive were classified as survivors.

Data Collection and Variables

Data were extracted retrospectively from hospital medical records and electronic health systems. The collected variables included:

- **Demographic characteristics:** age and sex
- **Clinical variables at admission:** vital signs such as heart rate, blood pressure, respiratory rate, oxygen saturation, and Glasgow Coma Scale
- **Laboratory parameters:** haemoglobin, total leukocyte count, platelet count, serum creatinine,

serum lactate, liver function tests, and arterial blood gas parameters

- **Comorbidities:** presence of chronic medical conditions including diabetes mellitus, hypertension, chronic kidney disease, chronic liver disease, and cardiovascular disease
- **Treatment-related variables:** need for vasopressors, mechanical ventilation, and intensive care unit admission
- **Severity scores:** Sequential Organ Failure Assessment (SOFA), Acute Physiology and Chronic Health Evaluation II (APACHE II), and National Early Warning Score 2 (NEWS2), calculated using standard definitions

Only variables available within the early phase of hospital admission were considered for model development to ensure clinical applicability.

Data Preprocessing

Data preprocessing was performed prior to model development. Continuous variables were inspected for outliers and implausible values. Missing data were handled using appropriate imputation strategies depending on the modeling approach. For conventional statistical models, missing continuous variables were imputed using median values, whereas tree-based machine learning models utilized intrinsic handling of missing values. Categorical variables were encoded numerically as required.

The dataset was assessed for class imbalance between survivors and non-survivors, and appropriate resampling techniques were applied during model training to reduce bias toward the majority class.

Model Development

Multiple predictive models were developed to estimate the probability of in-hospital mortality. These included:

- Logistic Regression
- Naïve Bayes
- K-Nearest Neighbors
- Support Vector Machine
- Decision Tree
- Random Forest
- Adaptive Boosting (AdaBoost)
- Extreme Gradient Boosting (XGBoost)

Model hyperparameters were optimized using grid search and cross-validation within the training dataset. Feature selection was based on clinical relevance and model-specific importance measures.

Model Training and Validation

The dataset was randomly divided into training and testing subsets. Models were trained exclusively on the training dataset and evaluated on the independent testing dataset to assess generalizability. Five-fold cross-validation was used during training to reduce overfitting and ensure robustness of performance estimates.

Performance Evaluation

Model performance was evaluated using multiple complementary metrics:

- Discrimination: assessed using the area under the receiver operating characteristic curve (AUROC)
- Calibration: evaluated using calibration plots and Brier score
- Classification performance: assessed using sensitivity, specificity, accuracy, and confusion matrices
- Clinical utility: evaluated using decision curve analysis to quantify net benefit across a range of threshold probabilities

The performance of machine learning models was directly compared with traditional severity scores (SOFA, APACHE II, and NEWS2).

Model Interpretability

To enhance interpretability, feature importance was examined for tree-based models. SHapley Additive exPlanations (SHAP) were used to quantify the contribution of individual predictors to model outputs and to provide both global and patient-level explanations of mortality risk.

Statistical Analysis

Descriptive statistics were used to summarize baseline characteristics of survivors and non-survivors. Continuous variables were expressed as mean with standard deviation or median with interquartile range, as appropriate. Categorical variables were summarized as frequencies and percentages. Statistical analyses and model development were performed using standard statistical and machine learning software packages.

RESULTS

Table 1. Baseline Characteristics of the Study Population According to In-Hospital Mortality

Variable	Survivors (n = 330)	Non-Survivors (n = 170)	p-value
Age (years), mean ± SD	55.1 ± 13.4	62.4 ± 12.6	<0.001
Male sex, n (%)	195 (59.1%)	111 (65.3%)	0.048
Diabetes mellitus, n (%)	138 (41.8%)	102 (60.0%)	0.002
Chronic kidney disease, n (%)	61 (18.5%)	66 (38.8%)	<0.001
ICU admission, n (%)	181 (54.8%)	140 (82.4%)	<0.001

Mechanical ventilation, n (%)	99 (30.0%)	113 (66.5%)	<0.001

Among the 500 patients included in the study, 170 (34%) died while 330 survived during hospitalization. Non-survivors were significantly older than survivors and had a higher proportion of males. Comorbid conditions such as diabetes mellitus and chronic kidney disease were more prevalent among non-survivors. The need for intensive care unit admission and mechanical ventilation was significantly higher in patients who died, indicating greater illness severity at presentation.

Table 2. Admission Clinical and Laboratory Parameters Among Survivors and Non-Survivors

Parameter	Survivors	Non-Survivors	p-value
Mean arterial pressure (mmHg)	76.8 ± 12.4	64.2 ± 14.1	<0.001
Serum lactate (mmol/L), median (IQR)	2.1 (1.6–3.2)	4.6 (3.1–6.8)	<0.001
Serum creatinine (mg/dL)	1.4 ± 0.8	2.6 ± 1.5	<0.001
Platelet count (×10⁹/L)	178 ± 64	112 ± 58	<0.001
Total leukocyte count (×10⁹/L)	13.6 ± 5.4	18.9 ± 7.1	<0.001

Non-survivors demonstrated significantly worse clinical and laboratory parameters at admission compared with survivors. They had lower mean arterial pressure and markedly higher serum lactate levels. Renal dysfunction, reflected by elevated serum creatinine, and hematological abnormalities, including thrombocytopenia and leukocytosis, were also significantly more common among patients who did not survive.

Table 3. Severity Scores Among Survivors and Non-Survivors

Score	Survivors	Non-Survivors	p-value
SOFA score, median (IQR)	5 (3–7)	10 (8–13)	<0.001
APACHE II score, mean ± SD	16.8 ± 6.2	26.4 ± 7.8	<0.001
NEWS2 ≥ 7, n (%)	96 (28.9)	122 (72.6)	<0.001

Severity scores at admission were substantially higher among non-survivors. Median SOFA scores and mean APACHE II scores were significantly elevated in patients who died during hospitalization. A significantly larger proportion of non-survivors had NEWS2 scores ≥7, highlighting the strong association between early severity scores and in-hospital mortality.

Table 4. Performance of Machine Learning Models for Predicting In-Hospital Mortality

Model	AUROC	Sensitivity	Specificity	Accuracy
Logistic Regression	0.81	72.4	79.2	76.9
Naïve Bayes	0.78	70.1	76.9	74.6
K-Nearest Neighbors	0.79	69.8	76.4	74.2
Support Vector Machine	0.82	72.9	79.8	77.1
Random Forest	0.84	75.3	82.0	79.1
AdaBoost	0.83	74.1	81.5	78.6
XGBoost	0.85	76.5	83.0	80.2

Machine learning models demonstrated variable predictive performance for in-hospital mortality. Traditional models such as logistic regression and Naïve Bayes showed moderate discrimination, whereas ensemble-based approaches performed better. The XGBoost model achieved the highest predictive accuracy, with superior AUROC, sensitivity, specificity, and overall accuracy compared with other machine learning algorithms.

Table 5. Comparison of Traditional Severity Scores and Best-Performing ML Model

Model / Score	AUROC	Brier Score
SOFA score	0.71	0.21
APACHE II	0.73	0.20
NEWS2	0.69	0.23
XGBoost model	**0.87**	**0.14**

When compared with traditional severity scoring systems, the XGBoost model demonstrated substantially better discrimination and calibration. The AUROC of the XGBoost model exceeded those of SOFA, APACHE II, and NEWS2 scores, and it also achieved a lower Brier score, indicating more accurate probability estimation of in-hospital mortality.

Table 6. Top Predictors of In-Hospital Mortality Identified by XGBoost Model

Rank	Predictor	Direction of Association
1	SOFA score	Higher scores ↑ mortality
2	Serum lactate	Higher values ↑ mortality
3	Mean arterial pressure	Lower values ↑ mortality
4	APACHE II score	Higher scores ↑ mortality
5	Serum creatinine	Higher values ↑ mortality

DISCUSSION

In this retrospective study, we evaluated the performance of multiple machine-learning models for predicting in-hospital mortality among patients with sepsis and compared their predictive ability with traditional severity scoring systems. Our findings demonstrate that machine-learning approaches, particularly ensemble models such as XGBoost, significantly outperformed conventional severity scores in terms of discrimination and calibration. These results reinforce the growing evidence supporting data-driven prognostic tools for risk stratification in sepsis.

The observed in-hospital mortality rate of approximately one-third of patients in our cohort is consistent with previously reported mortality rates in hospitalized and critically ill sepsis populations, particularly in resource-constrained settings. Rudd et al. [1] highlighted substantial variability in sepsis mortality across regions, with higher rates reported in low- and middle-income countries. The demographic and clinical differences observed between survivors and non-survivors in our study—such as advanced age, higher prevalence of comorbidities, and greater need for ICU admission and mechanical ventilation—are well aligned with prior observational studies on sepsis outcomes (Seymour et al. [11], Vincent et al. [12]).

Laboratory and physiological derangements at admission were strongly associated with mortality in our cohort. Elevated serum lactate, renal dysfunction, thrombocytopenia, and hypotension were significantly more common among non-survivors. Lactate, in particular, emerged as the most influential predictor in the XGBoost model. This finding is consistent with prior studies demonstrating the prognostic value of hyperlactatemia as a marker of tissue hypoperfusion and metabolic stress in sepsis (Singer et al. [2], Vincent et al. [12]). Serial and admission lactate measurements have repeatedly been shown to correlate with mortality risk, supporting its prominence in predictive models.

Traditional severity scores such as SOFA, APACHE II, and NEWS2 were significantly higher among non-survivors, confirming their utility in identifying high-risk patients. However, their discriminatory performance in our study was modest when compared with machine-learning models. Previous studies have reported similar limitations of conventional scores, particularly their reliance on static measurements and linear assumptions (Knaus et al. [3], Kim et al. [4]). These limitations reduce their ability to capture complex nonlinear interactions and evolving physiological patterns characteristic of sepsis.

Machine-learning models demonstrated superior predictive performance, with ensemble methods consistently outperforming single classifiers. Among these, XGBoost achieved the highest AUROC and lowest Brier score, indicating both strong discrimination and reliable probability estimation. These findings are in agreement with prior retrospective studies using electronic health record data, where gradient-boosting methods outperformed logistic regression and traditional scores for mortality prediction in sepsis (Li et al. [13], Liang et al. [7]). The ability of XGBoost to handle missing data, model nonlinear relationships, and reduce overfitting through regularization likely contributed to its superior performance.

The comparison between machine-learning models and traditional severity scores in our study highlights the potential clinical value of ML-based prognostic tools. While SOFA and APACHE II remain useful for bedside assessment and benchmarking, their predictive accuracy was inferior to that of the XGBoost model. Similar observations have been reported in large ICU datasets, including MIMIC-based studies, where ML models demonstrated improved mortality prediction across diverse patient populations (Johnson et al. [5], Desautels et al. [8]). Importantly, improved discrimination alone is insufficient; the lower Brier score observed with XGBoost suggests more accurate risk estimation, which is critical for clinical decision-making.

Interpretability remains a key challenge in the adoption of machine-learning models in critical care. In this study, feature importance and SHAP-based explanations identified clinically intuitive predictors such as lactate, SOFA score, renal dysfunction, hypotension, vasopressor requirement, thrombocytopenia, and age as major contributors to mortality risk. The alignment of these predictors with established clinical knowledge enhances trust in the model and supports its potential clinical applicability. Previous studies have emphasized that explainable machine-learning approaches improve clinician acceptance and facilitate integration into decision-support systems (Lundberg et al. [14], Patil AR [15], Van Calster et al. [10]).

Despite the strengths of our study, several limitations should be acknowledged. The retrospective design introduces the possibility of selection bias and unmeasured confounding. The data were derived from a single tertiary-care center, which may limit generalizability to other healthcare settings. Additionally, while internal validation was performed, external validation using independent

datasets is necessary before clinical implementation. These limitations are common to retrospective ML studies and have been highlighted in prior methodological reviews (Steyerberg et al. [16], Collins et al. [17], Patil AR et al. [18], Bagul DY et al . [19, 20]).

LIMITATIONS OF THE STUDY

This study was limited by its retrospective design, which may be subject to selection bias and unmeasured confounding. As the analysis was conducted at a single tertiary care center, the generalizability of the findings to other settings may be limited. The use of routinely collected clinical data may have introduced missing or inaccurate measurements despite appropriate preprocessing. Additionally, external validation using independent datasets was not performed and is necessary before clinical implementation of the models.

CONCLUSION

In conclusion, this study demonstrated that machine-learning models, particularly gradient-boosting approaches, provided superior prediction of in-hospital mortality in patients with sepsis compared with traditional severity scoring systems. By effectively utilizing routinely collected clinical data and incorporating model explainability techniques, these approaches enabled improved early risk stratification and supported more individualized clinical decision-making. Further research should focus on external validation, prospective assessment, and seamless integration of machine-learning based models into clinical workflows to determine their real-world impact on sepsis management and patient outcomes.

REFERENCES

1. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. **Lancet**. 2020;395(10219):200–11.
2. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). **JAMA**. 2016;315(8):801–10.
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. **Crit Care Med**. 1985;13(10):818–29.
4. Kim WY, Huh JW, Hong SB, Lim CM, Koh Y. Predicting mortality in sepsis patients using SOFA score and clinical variables. **J Crit Care**. 2020;55:186–92.
5. Johnson AEW, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. **Sci Data**. 2016;3:160035.
6. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? **PLoS One**. 2017;12(4):e0174944.
7. Liang Y, Li X, Chen T, Zhang J, Chen M, Chen J, et al. Predicting mortality in sepsis patients using machine learning models. **BMC Med Inform Decis Mak**. 2020;20:251.
8. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. **JMIR Med Inform**. 2016;4(3):e28.
9. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence for early prediction of sepsis using electronic health records. **Nat Commun**. 2021;12:711.
10. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. **BMC Med**. 2019;17:230.
11. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). **JAMA**. 2016;315(8):762–74.
12. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. **Intensive Care Med**. 1996;22(7):707–10.
13. Li X, Xu X, Xie F, et al. A machine learning–based model for predicting in-hospital mortality in patients with sepsis using electronic health records. **BMC Med Inform Decis Mak**. 2022;22:84.
14. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. **Adv Neural Inf Process Syst**. 2017;30:4765–74.
15. Patil AR, Bharate PB, Junaid M. Enhancing myocardial infarction diagnosis with efficient machine learning techniques through combination of correlation and variance threshold feature selection. **Int J Life Sci Biotechnol Pharma Res**. 2023;12(4):172–180.
16. Steyerberg EW, Harrell FE Jr, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. **J Clin Epidemiol**. 2001;54(8):774–81.
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. **Ann Intern Med**. 2015;162(1):55–63.
18. Patil AR, Bharate PB, Junaid M. Machine learning based myocardial infarction risk stratification as a diagnostic aid for remote areas with limited medical

resources. **Journal of Cardiovascular Disease Research**. 2023;14(10):790–800.

19. Bagul DY, Bharate PB, Sahasrabuddhe A. Use of robust machine learning approach in prediction of stroke. J Cardiovasc Dis Res. 2023;14(11):256-260

20. Bagul DY, Bharate PB, Sahasrabuddhe A. Prediction of stroke with extreme gradient boosting in machine learning model. Int J Life Sci Biotechnol Pharma Res. 2024;13(7):249-253.