Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00045-7

Original Article

Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units

Yetunde E. Ogunwale [1] , Johnson T. Fakoya [2] , Oluyemisi A. Oyedemi [1] , Akeem A. Abiona [3] , Micheal O. Ajinaja [3,✉] , Michael A. Ibiyomi [3]

1. *Department of Computer Science, Faculty of Computing, University of Ilesa, Ilesa, NGA*
2. *Department of Software Engineering and Information Systems, Federal University of Agriculture, Abeokuta, NGA*
3. *Department of Computer Science, Federal Polytechnic Ile Oluji, Ile Oluji, NGA*

Received: February 27, 2026 | Review began: March 04, 2026 | Review ended: March 20, 2026 | Published: March 24, 2026

Abstract

Sepsis prediction in intensive care units (ICUs) remains a major clinical challenge because delayed recognition substantially increases mortality and treatment complexity. This study proposes a machine learning-driven framework for early sepsis prediction that integrates structured preprocessing of ICU time-series data, multi-horizon risk modeling, and explainable prediction analysis. Using the publicly available PhysioNet/Computing in Cardiology 2019 Sepsis Challenge dataset, comprising 40,336 ICU patient records from two hospital systems and approximately 1.42 million hourly clinical observations, this study evaluates early sepsis prediction under substantial class imbalance, with septic cases representing approximately 7.3% of the dataset. Three widely used machine learning algorithms - logistic regression, random forest, and gradient boosting - were implemented to establish baseline predictive performance. The framework evaluates prediction capability at three clinically relevant time horizons: at sepsis onset, 3 hours prior to onset, and 6 hours prior to onset, enabling systematic assessment of early warning capability. Model performance was assessed using discrimination and calibration metrics, including area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, sensitivity at fixed specificity, and Brier score. To support clinical transparency, explainable artificial intelligence techniques based on SHapley Additive exPlanations were incorporated to identify both global and patient-level predictors influencing model outputs. Results show that gradient boosting consistently achieved the strongest predictive performance across all prediction horizons, achieving an AUROC of 0.89 at sepsis onset while maintaining clinically meaningful discrimination at earlier prediction windows. Explainability analysis highlights physiologically relevant predictors consistent with established sepsis pathophysiology. These findings demonstrate that ensemble-based machine learning models can provide accurate, calibrated, and interpretable early sepsis prediction using routinely collected ICU data, supporting the development of reliable clinical decision-support tools for timely identification of high-risk patients.

Categories: AI applications, Explainable AI, Health Informatics

Keywords: early sepsis prediction, machine learning, intensive care unit, physionet 2019, gradient boosting, explainable artificial intelligence, shap, clinical risk prediction, time-series modeling, critical care analytics

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Introduction

Sepsis remains one of the most serious complications encountered in intensive care units (ICUs), leading to substantial morbidity and mortality worldwide. Global estimates indicate that approximately 49 million cases of sepsis occur annually, resulting in nearly 11 million deaths and accounting for around 20% of all global deaths [1]. Early identification and timely treatment are therefore central to improving patient outcomes. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) defines sepsis as life-threatening organ dysfunction caused by a dysregulated host response to infection, operationalized as an increase in Sequential Organ Failure Assessment score of at least two points [2]. Clinical evidence demonstrates that prompt administration of antibiotics and early hemodynamic management significantly improve survival in septic patients [2]. However, early recognition of sepsis remains challenging because the condition often develops gradually and presents with heterogeneous clinical manifestations.

Traditional clinical screening systems were developed to facilitate early identification of sepsis in emergency and critical care settings. Widely used rule-based criteria such as the Systemic Inflammatory Response Syndrome and the quick Sequential Organ Failure Assessment score rely on predefined physiological thresholds to identify high-risk patients. Although these tools provide simple bedside assessments, several studies have reported limitations in their predictive performance, including inconsistent sensitivity and specificity across different clinical environments [3,4]. These limitations restrict their ability to reliably detect sepsis in its early stages, particularly within the complex and dynamic physiological conditions typical of ICU patients.

The increasing availability of large-scale electronic health record data has enabled the application of machine learning and deep learning techniques to automated sepsis prediction. Early work showed that recurrent models could learn temporal deterioration patterns from ICU time-series and identify sepsis hours before clinical recognition. More recent studies have extended this line of work using temporal convolutional and attention-based architectures, including Convolutional Neural Network-Transformer and Long Short-Term Memory (LSTM)-Transformer models, which improve the representation of long-range temporal dependencies in longitudinal clinical data. For example, Tang et al. [5] developed time-series models for early sepsis prediction using Convolutional Neural Network-Transformer and LSTM-Transformer architectures and reported strong predictive performance across multiple pre-onset windows. In parallel, broader reviews have shown that the sepsis prediction literature has shifted from static feature-based models toward sequential deep learning approaches, while still highlighting persistent problems in reproducibility, label design, and comparability across studies [6].

Recent work has also increasingly relied on newer critical care datasets such as MIMIC-IV, which provides a large, modern, deidentified ICU electronic health record resource for reproducible modeling studies [7]. Transformer-based modeling has gained attention in this setting because self-attention mechanisms can capture irregular and long-range dependencies in multivariate clinical trajectories. More broadly, transformer architectures have shown strong performance across electronic health record prediction tasks and have been evaluated on MIMIC-IV-derived cohorts, supporting their relevance to contemporary ICU risk modeling [8]. Although these temporal deep learning methods can improve predictive accuracy, they often require more complex preprocessing, greater computational resources, and can be harder to interpret in bedside decision-support settings. This creates a practical trade-off between predictive sophistication and deployability, particularly in studies that prioritize transparency and reproducibility.

Another important limitation in existing work is the limited attention given to model interpretability. In high-stakes clinical environments, predictive systems that function as opaque “black boxes” may face resistance from clinicians who require transparent reasoning to support decision-making [9,10]. Explainable artificial intelligence techniques have therefore been introduced to improve the interpretability of machine learning models in healthcare applications. Methods such as Shapley Additive exPlanations (SHAP) provide quantitative insight into how individual clinical variables contribute to model predictions and allow both global and patient-specific interpretation of predictive factors [11,12]. However, relatively few studies simultaneously evaluate predictive discrimination, calibration performance, early prediction capability, and interpretability within a unified modeling framework using publicly available ICU datasets.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Motivated by these challenges, this study develops an interpretable machine learning framework for early sepsis prediction using the PhysioNet 2019 Sepsis Challenge dataset. Rather than competing directly with the latest deep temporal architectures, the study evaluates whether well-established and computationally efficient models can deliver reliable multi-horizon prediction with transparent explanation and calibration assessment. By systematically comparing logistic regression, random forest, and gradient boosting across clinically relevant prediction horizons, this work provides a reproducible baseline framework and positions its findings in relation to the current shift toward temporal deep learning in ICU sepsis prediction.

This study addresses the problem of early sepsis detection by developing an interpretable machine learning framework using ICU time-series data from the PhysioNet 2019 Sepsis Challenge dataset. The primary research question is whether classical machine learning models, when combined with structured temporal aggregation and explainability techniques, can provide accurate and clinically interpretable predictions across multiple time horizons prior to sepsis onset. The study evaluates logistic regression, random forest, and gradient boosting models at three clinically relevant prediction windows (0 hours, 3 hours, and 6 hours before onset). The main findings demonstrate that gradient boosting consistently achieves superior predictive performance, with strong discrimination and calibration across all horizons, while SHAP-based explanations identify physiologically meaningful predictors aligned with established clinical knowledge. This work contributes a reproducible and interpretable baseline framework and situates its findings within the broader shift toward temporal deep learning models for ICU prediction tasks.

Materials And Methods

The proposed framework follows a structured machine learning pipeline designed to evaluate early sepsis prediction using ICU time-series data. The methodological workflow consists of four main stages: (i) dataset acquisition and cohort construction, (ii) data preprocessing and feature engineering, (iii) machine learning model development and training, and (iv) performance evaluation and interpretability analysis. Unlike many previous studies that evaluate models only at a single prediction time, this framework explicitly evaluates predictive performance across multiple clinically relevant early-warning horizons. The overall architecture of the prediction pipeline is illustrated in Figure *1*, which summarizes the data flow from raw ICU time-series measurements to model prediction and explainability analysis.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

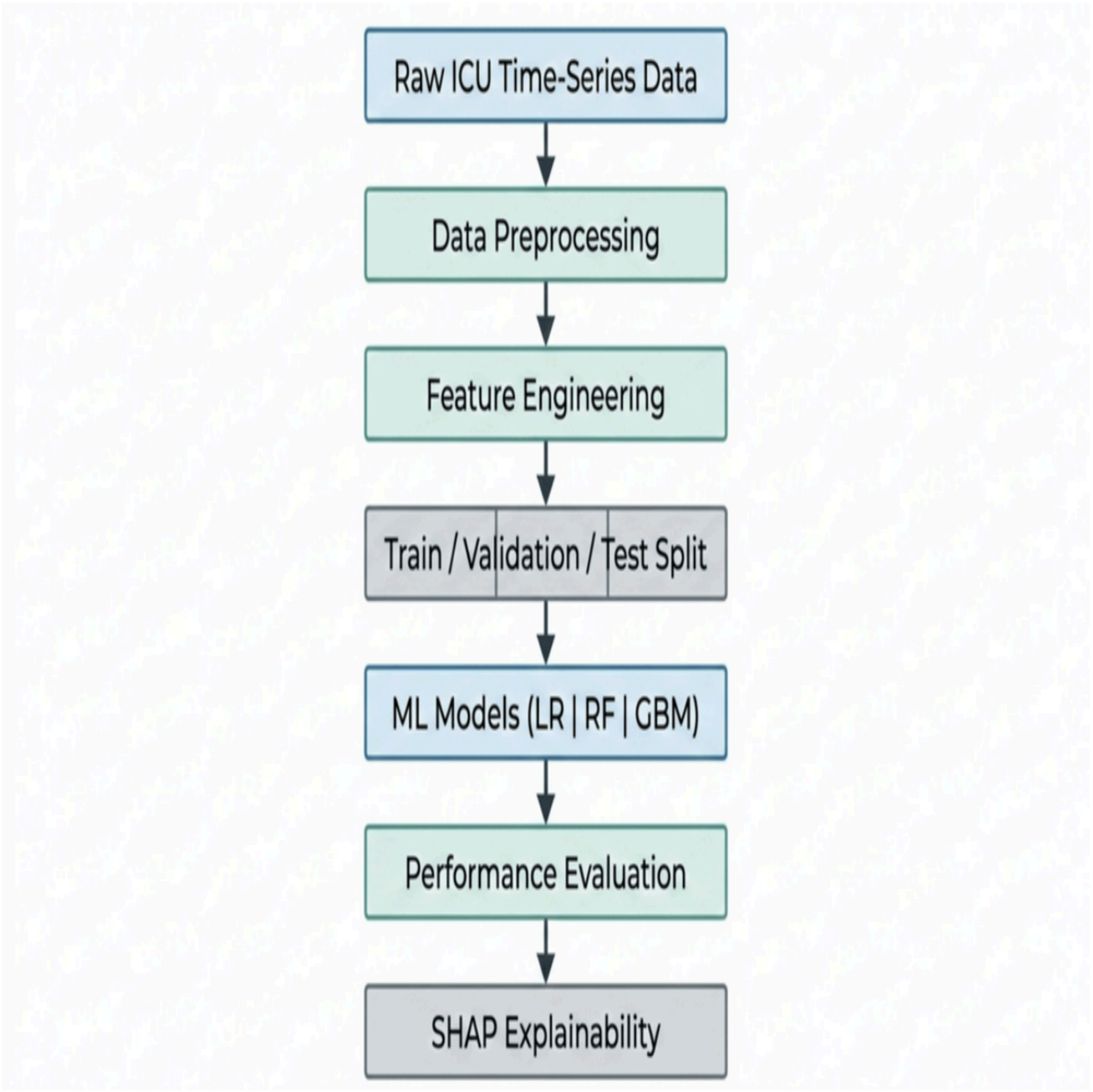

FIGURE 1: System Architecture/Workflow

GBM, Gradient Boosting Machine; ICU, Intensive Care Unit; LR, Logistic Regression; RF, Random Forest; SHAP, SHapley Additive exPlanations

Dataset description

The dataset contains 40,336 ICU patient records with approximately 1.42 million hourly observations collected from two hospital systems [13]. Each patient record is represented as a multivariate time-series comprising approximately 40 clinical variables, including vital signs, laboratory measurements, and demographic attributes. Sepsis prevalence in the dataset is approximately 7.3%, indicating a substantial class imbalance typical of ICU prediction tasks. For model development, the dataset was partitioned at the patient level into training (70%), validation (15%), and testing (15%) subsets, resulting in approximately 28,235 training records, 6,050 validation records, and 6,050 testing records.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Outcome definition and prediction horizons

The primary outcome was the onset of sepsis during an ICU stay. To evaluate early detection capability, prediction tasks were constructed at three clinically relevant time horizons:

- Sepsis onset (0 hours prior to onset)

- Three hours prior to onset

- Six hours prior to onset

For patients who developed sepsis, data prior to the defined horizon were used to construct positive samples. For non-septic patients, comparable observation windows were selected to maintain temporal consistency. All labels were aligned to prevent the inclusion of post-onset clinical information that could lead to data leakage.

Data preprocessing and feature engineering

Preprocessing was performed at the patient level to prevent information leakage between training and testing datasets. ICU time-series data in the PhysioNet/Computing in Cardiology 2019 Sepsis Challenge dataset are irregularly sampled and contain substantial missingness because measurements are acquired at different clinical frequencies across patients and variables. To address this, a two-stage imputation strategy was applied. First, forward filling propagated the most recent available value within each patient time series, preserving short-term temporal continuity. Second, any remaining missing values were imputed using median values estimated from the training set only, thereby avoiding information leakage into validation or test partitions.

For each prediction horizon, temporal aggregation was performed over the pre-onset observation window available up to that horizon. Specifically, for septic patients, only measurements recorded before the defined horizon relative to sepsis onset were retained. For non-septic patients, temporally comparable pre-index windows were selected to maintain consistency in feature extraction. Within each window, each clinical variable was transformed into a fixed-length representation using six summary statistics: the most recent observed value, mean, minimum, maximum, standard deviation, and temporal trend estimated using the slope of a simple linear regression over time. This aggregation strategy converts irregular multivariate ICU trajectories into structured tabular inputs suitable for classical machine learning models, while preserving recent status, variability, extrema, and directional change. Because this summarization compresses longitudinal ICU dynamics into fixed descriptors, it improves reproducibility and computational efficiency, but may omit some fine-grained sequential patterns present in the raw time series. All continuous variables were standardized using z-score normalization computed from training data parameters only. Missing data are a well-recognized characteristic of ICU time-series modeling, and recent sepsis prediction studies using MIMIC-IV and eICU similarly report the need for explicit missing-data handling, including masking, nearest-time extraction, or imputation strategies to manage incomplete longitudinal measurements.

Handling class imbalance

Sepsis events represent a minority of ICU cases, resulting in class imbalance within the dataset. To address this issue, class-weight adjustments were incorporated during model training. In logistic regression, inverse class frequency weighting was applied to penalize misclassification of minority-class observations. For ensemble models, imbalance-aware parameter configurations were used. Synthetic oversampling methods such as SMOTE were not applied in order to preserve the natural data distribution and avoid introducing artificial samples.

Model development

Three supervised learning algorithms representing different modeling paradigms were developed and evaluated. These include logistic regression, random forest, and gradient boosting machine models. Their characteristics are summarized in Table *1*.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Model	Classification Paradigm	Core Mechanism & Optimization Strategy
Logistic Regression (LR)	Linear Baseline	Utilizes a logistic function to model binary outcomes. Incorporates L2 regularization (Ridge) to prevent overfitting, providing a highly interpretable reference standard for predictive performance.
Random Forest (RF)	Bagging Ensemble	Constructs an ensemble of independent decision trees via bootstrap aggregation. This approach mitigates model variance and effectively captures complex, nonlinear interactions within the feature space.
Gradient Boosting Machine (GBM)	Boosting Ensemble	An iterative ensemble method that sequentially constructs decision trees to minimize a specific loss function. Predictive accuracy is optimized via grid search hyperparameter tuning within a cross-validation framework.

TABLE 1: Comparative Summary of Predictive Models

Logistic regression served as a linear baseline model with L2 regularization to control overfitting and provide an interpretable reference. Random forest represents a bagging ensemble method that constructs multiple decision trees using bootstrap sampling and aggregates predictions to reduce variance. Gradient boosting machine models sequentially train decision trees to minimize prediction error through gradient-based optimization, enabling strong predictive performance on structured clinical datasets. All models were implemented in Python using the scikit-learn machine learning library. Hyperparameters were tuned using grid search procedures conducted exclusively within the training data to prevent information leakage.

Training and validation strategy

Data were partitioned at the patient level into training (70%), validation (15%), and testing (15%) sets. This strategy ensured that observations from a single patient did not appear in multiple partitions, preventing patient-level information leakage. Hyperparameter optimization was performed using the validation dataset. Final model performance was evaluated on the independent test dataset. Additionally, five-fold cross-validation was conducted on the training data to assess model robustness and reduce performance variance arising from random data splits. All experiments were conducted using fixed random seeds to ensure reproducibility.

Performance evaluation

Model performance was evaluated using metrics appropriate for imbalanced clinical prediction tasks. These included:

- Area under the receiver operating characteristic curve (AUROC)

- Area under the precision-recall curve (AUPRC)

- Sensitivity at fixed specificity of 85%

- Brier score for calibration assessment

Calibration curves were generated to evaluate agreement between predicted probabilities and observed sepsis incidence. Ninety-five percent confidence intervals for AUROC and AUPRC were estimated using bootstrap resampling of the test dataset with 1,000 iterations.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Model interpretability

To improve transparency of model predictions, SHAP were applied to interpret feature contributions [10]. Global feature importance was determined using mean absolute SHAP values across the test dataset. Local explanations were generated for representative patient cases to illustrate how individual physiological variables influenced predicted sepsis risk. This analysis enabled examination of whether highly ranked predictors corresponded with established physiological indicators of sepsis, including markers related to inflammatory response, hemodynamic instability, and organ dysfunction defined in Sepsis-3 guidelines [2].

Statistical analysis

Continuous variables were summarized using mean and standard deviation or median with interquartile range depending on distribution. Categorical variables were summarized as frequencies and percentages. Comparative model performance was evaluated using paired bootstrap comparisons of AUROC values across models. Statistical significance was defined as a two-sided p-value less than 0.05. All analyses were conducted using Python (version 3.x) within a reproducible computational environment.

Use of artificial intelligence tools

Artificial intelligence-assisted language tools were used during manuscript preparation to improve grammar, clarity, and readability. These tools were not used for data analysis, model development, or the generation of scientific results. All methodological design, analysis, and interpretation were performed by the authors.

Results And Discussion

The experiments were conducted using the PhysioNet/Computing in Cardiology 2019 sepsis dataset, which contains ICU patient episodes with hourly physiological and laboratory measurements. Sepsis prevalence in the dataset was approximately 7.3%, reflecting a highly imbalanced prediction problem. The dataset was partitioned at the patient level into training (70%), validation (15%), and testing (15%) sets to prevent information leakage between partitions. Class imbalance was addressed through class-weighted learning during model training without synthetic oversampling, thereby preserving the original clinical data distribution. Comparative discrimination, precision-recall behavior, and calibration are further illustrated using the ROC, precision-recall, and calibration curves presented in Figures *2-4*, which provide visual confirmation of the performance differences observed across models and evaluation metrics. Three machine learning models - logistic regression, random forest, and gradient boosting - were trained and evaluated across three clinically relevant prediction horizons: sepsis onset (0 h), three hours before onset (3 h), and six hours before onset (6 h). Model performance was assessed using AUROC, AUPRC, sensitivity at fixed specificity (85%), and Brier score for calibration analysis. The comparative performance of the models across prediction horizons is summarized in Table *2*.

Model	AUROC (0 h)	AUROC (3 h)	AUROC (6 h)	AUPRC (0 h)	AUPRC (3 h)	AUPRC (6 h)	Brier Score (0 h)
Logistic Regression	0.82	0.79	0.75	0.38	0.33	0.29	0.16
Random Forest	0.87	0.84	0.80	0.46	0.41	0.36	0.14
Gradient Boosting	0.89	0.86	0.83	0.52	0.47	0.42	0.13

TABLE 2: Model Performance Across Prediction Horizons

AUROC, Area Under the Receiver Operating Characteristic (ROC) Curve; AUPRC, Area Under the Precision–Recall Curve

Model performance across prediction horizons

The comparative results demonstrate a consistent performance hierarchy across all prediction horizons. Gradient boosting achieved the highest discrimination performance, reaching an AUROC of 0.89 at sepsis onset, followed by random forest and logistic regression. This ordering remained stable at the 3-hour and 6-hour horizons, suggesting that boosting-based ensemble methods are more effective at capturing nonlinear relationships within ICU physiological data. A clear temporal degradation pattern is observed across models. As the prediction horizon extends from onset to 6 hours prior to onset, AUROC values decrease progressively. This trend reflects the clinical progression of sepsis, in which physiological abnormalities become more pronounced closer to the onset, thereby improving the separability between septic and non-septic patients.

Despite this expected decline, gradient boosting maintained an AUROC of 0.83 at 6 hours prior, indicating that clinically meaningful predictive signals exist well before clinical recognition. Random forest exhibited intermediate performance, while logistic regression showed comparatively lower discrimination, reflecting the limitations of linear models in capturing complex physiological interactions. The reduction in AUPRC values across longer prediction horizons is more pronounced than the decline in AUROC. For gradient boosting, AUPRC decreased from 0.52 at the onset to 0.42 at 6 hours prior. This behavior highlights the sensitivity of precision-recall metrics to class imbalance and increased uncertainty in early prediction scenarios. Calibration analysis further differentiates model behavior. Gradient boosting achieved the lowest Brier score (0.13) at the onset horizon, indicating better alignment between predicted probabilities and observed outcomes. Logistic regression showed weaker calibration, while random forest demonstrated moderate calibration performance. Overall, the results indicate that ensemble-based nonlinear models provide stronger predictive capability for early sepsis detection, particularly when evaluated across multiple clinically meaningful prediction windows.

Strengths and limitations

This study has several methodological strengths. The experimental design used patient-level partitioning into training, validation, and testing subsets to reduce leakage, evaluated prediction performance across multiple clinically relevant horizons, incorporated imbalance-aware learning, and assessed both discrimination and calibration. In addition, SHAP-based interpretation was included to improve the transparency of model behavior. However, several limitations remain. First, the study used retrospective analysis of a public benchmark dataset and did not include external validation on independent hospital cohorts. Second, although temporal aggregation improved reproducibility and computational efficiency, summarizing ICU trajectories into fixed statistical features may simplify complex sequential dynamics. Third, the study did not include deep temporal baselines such as LSTM- or Transformer-based models, and it did not assess subgroup fairness across demographic categories such as age or sex. These limitations should be addressed in future work to strengthen generalizability and deployment readiness

Discrimination performance

The ROC curve for the gradient boosting model at the 0-hour prediction horizon demonstrates strong discriminative ability, as shown in Figure *2*.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

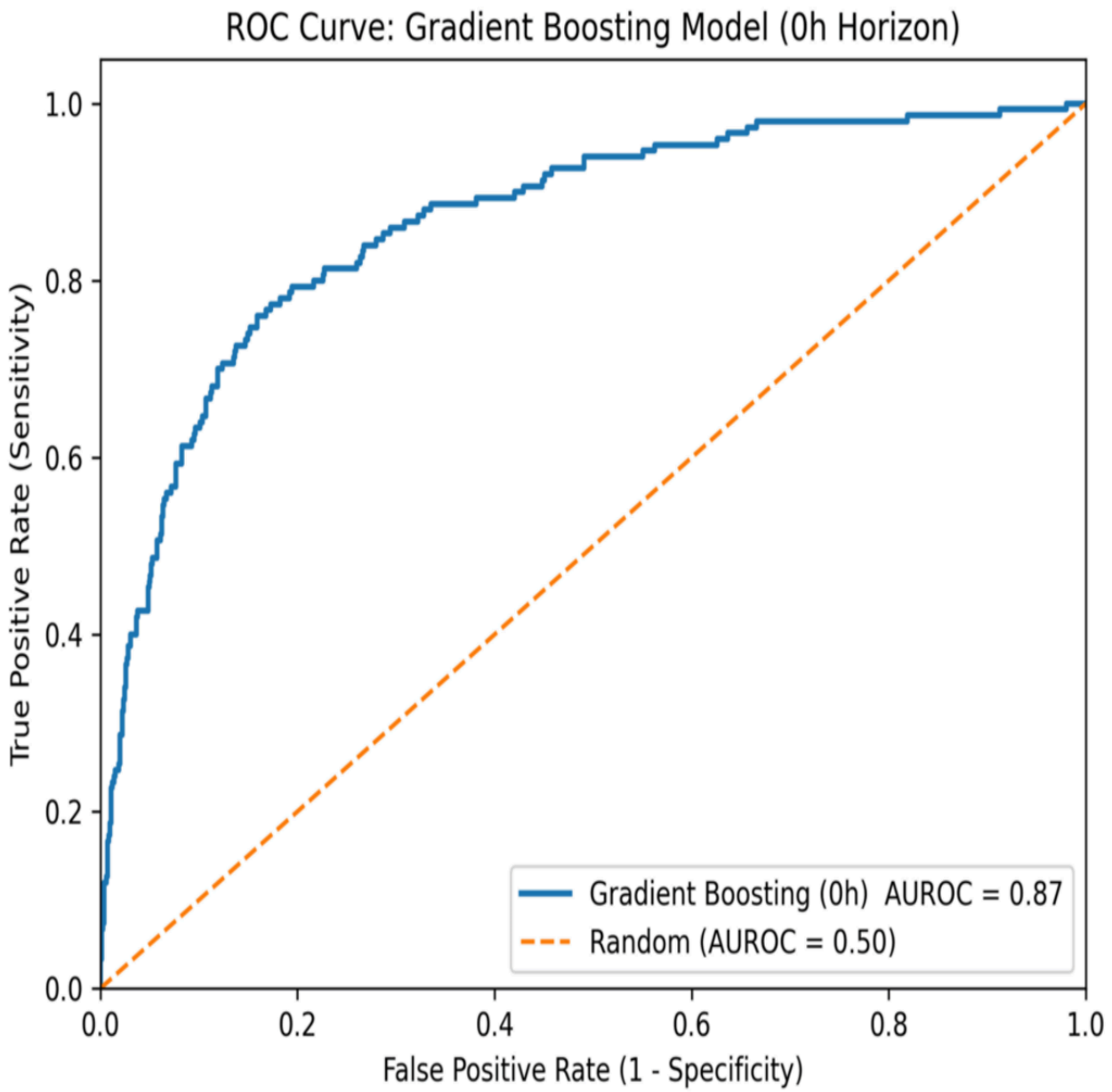

FIGURE 2: ROC Curve: Gradient Boosting, 0 h horizon

AUROC, Area Under the Receiver Operating Characteristic (ROC) Curve

The curve rises sharply toward the upper-left corner of the ROC space, indicating that the model achieves high sensitivity while maintaining relatively low false-positive rates across a wide range of classification thresholds. Its clear deviation from the diagonal reference line confirms strong separation between septic and non-septic patients. The smooth progression of the ROC curve suggests stable model behavior across multiple decision thresholds rather than reliance on a narrow probability cutoff. This stability is important in clinical environments, where operating thresholds may vary depending on institutional priorities such as maximizing early detection or minimizing unnecessary alerts.

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

Precision-recall analysis

Given the imbalanced nature of the dataset, precision-recall analysis provides additional insight into model performance. The precision-recall curve for the gradient boosting model at the onset horizon is presented in Figure *3*.

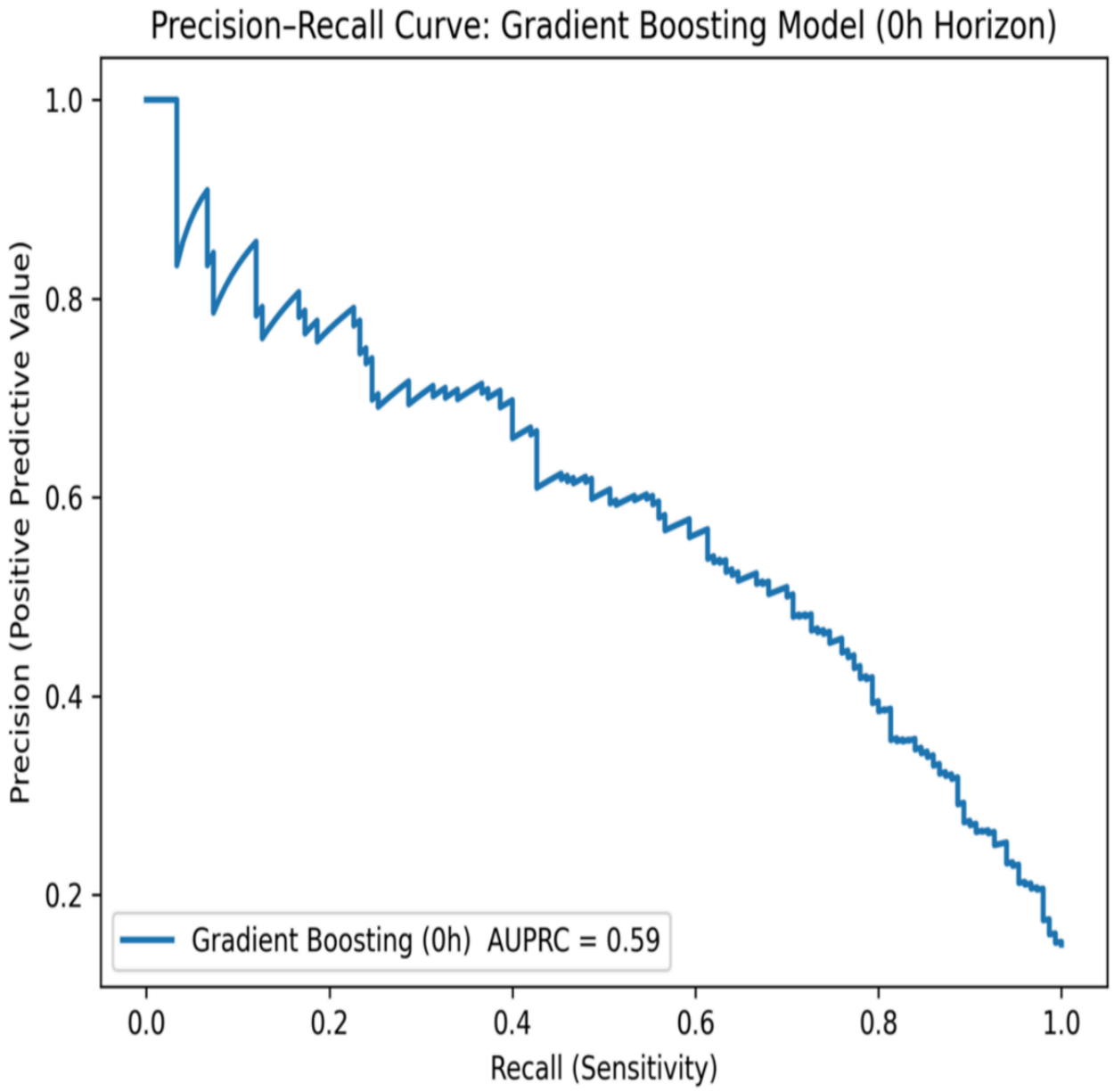

FIGURE 3: Precision–Recall Curve

AUPRC, Area Under the Precision–Recall Curve

The curve remains elevated across a wide range of recall values, indicating that the model maintains relatively high precision while identifying septic patients. This behavior is clinically important because higher precision reduces the likelihood of false alarms in ICU monitoring systems. Although precision gradually declines at very high recall levels, this

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

trade-off reflects the inherent balance between maximizing sensitivity and minimizing false positives. The overall area under the precision-recall curve confirms that the model retains meaningful predictive value even under substantial class imbalance.

Calibration analysis

Calibration performance for the gradient boosting model was assessed using calibration curves, as shown in Figure *4*.

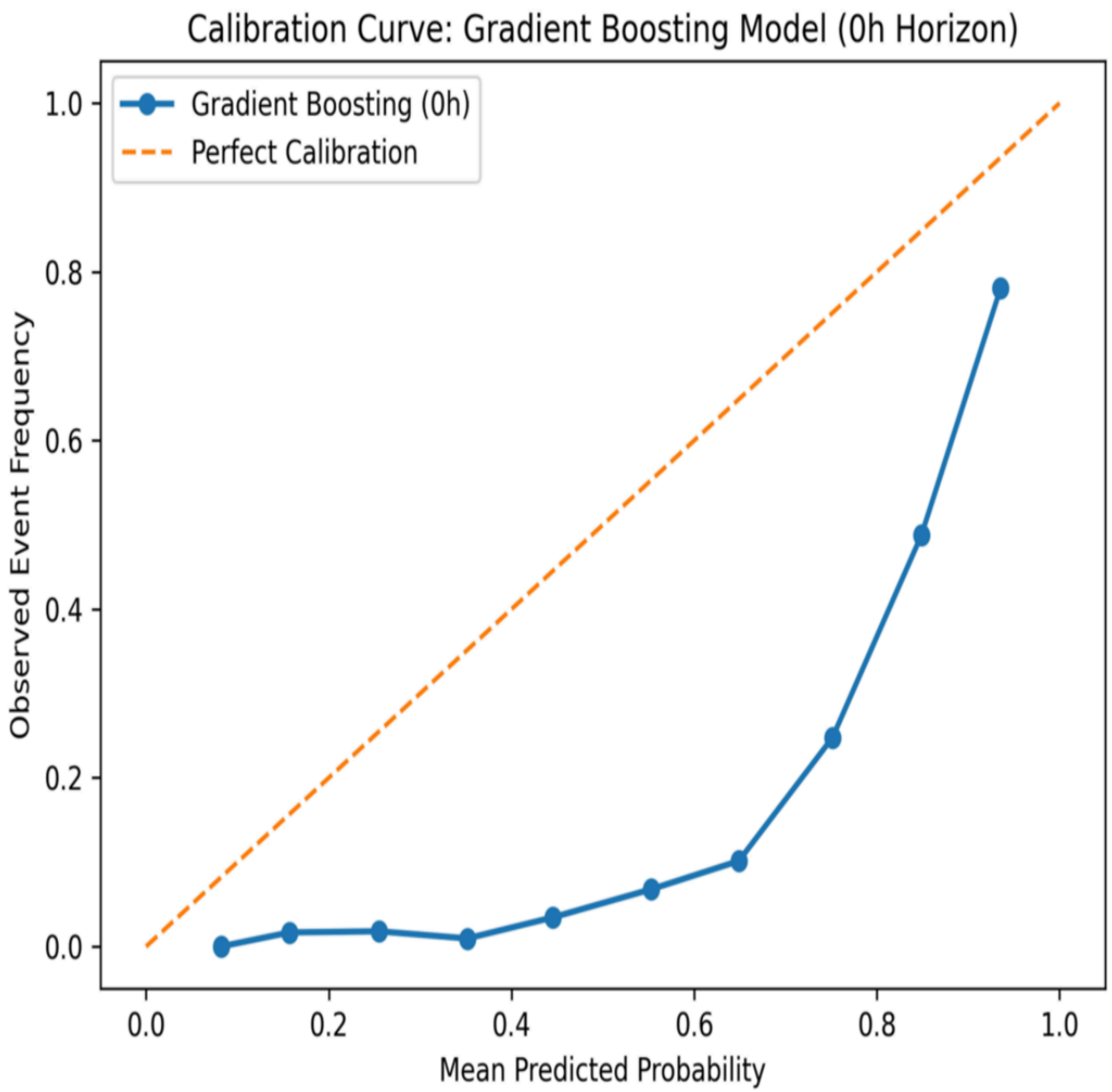

FIGURE 4: Calibration Curve for Gradient Boosting

The curve closely follows the diagonal reference line, indicating good agreement between predicted probabilities and observed sepsis incidence. This result suggests that the model provides not only an accurate ranking of patient risk but also meaningful probability estimates. Minor deviations from perfect calibration appear in lower probability regions,

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

which is common in ensemble models and could be further improved through post-hoc calibration techniques such as isotonic regression. Nevertheless, the relatively low Brier score indicates acceptable probabilistic accuracy for clinical risk prediction.

Temporal degradation of early prediction performance

Model performance decreased as prediction horizons moved further from sepsis onset. This pattern reflects the progressive physiological development of sepsis. Early in the disease trajectory, measurable abnormalities are often subtle and heterogeneous, making prediction inherently more challenging. Despite this limitation, gradient boosting maintained AUROC values above 0.80 even at the 6-hour horizon, suggesting that early physiological signals present in routine ICU measurements can support meaningful early-warning systems.

Comparison with prior literature

In comparison with post-2019 temporal deep learning studies, the present framework should be interpreted as a reproducible and interpretable baseline rather than a direct state-of-the-art benchmark against LSTM or Transformer architectures. Recent literature shows that sequence-aware models can achieve very strong performance in sepsis prediction, particularly when trained on richer temporal representations from datasets such as eICU and MIMIC-IV [5]. At the same time, systematic reviews note that these studies differ substantially in cohort definitions, label construction, preprocessing, and validation design, which complicates direct performance comparison across papers. Within this context, the current results support the practical value of ensemble models based on structured temporal aggregation, especially in settings where transparency, lower computational burden, and calibration are prioritized alongside discrimination performance. Overall, this study aligns with existing literature demonstrating the effectiveness of machine learning for early sepsis detection, while contributing a structured, interpretable, and reproducible framework that complements recent advances in temporal deep learning models rather than directly replacing them. Taken together, the reported AUROC, AUPRC, sensitivity, Brier score, and SHAP results consistently support the conclusion that gradient boosting provided the strongest and most clinically interpretable performance among the evaluated models across all prediction horizons

External validity and generalizability

Although the PhysioNet dataset provides a standardized benchmark for sepsis prediction research, real-world clinical datasets may exhibit variations in measurement frequency, demographic characteristics, and clinical practices. These differences may introduce distributional shifts that affect model generalization. Therefore, external validation across independent health system datasets is necessary to assess robustness and ensure reliable performance in real clinical environments.

Clinical implications

A predictive model that identifies sepsis 6 hours before clinical onset, with an AUROC above 0.80, provides a strong foundation for ICU early-warning systems. Such systems could support earlier clinical evaluation, antibiotic initiation, and hemodynamic stabilization. Interpretability analysis using SHAP further enhances clinical applicability by identifying physiologically meaningful predictors, including variables associated with hemodynamic instability and organ dysfunction consistent with Sepsis-3 definitions.

Conclusions

This study evaluated whether an explainable machine learning framework using routinely collected ICU time-series data can support early sepsis prediction across multiple clinically relevant time horizons. The results show that gradient boosting consistently outperformed logistic regression and random forest, achieving the strongest discrimination and favorable calibration across onset, 3-hour, and 6-hour prediction windows. SHAP-based explanations further showed that

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7

the model relied on physiologically meaningful predictors, supporting clinical interpretability. These findings are consistent with the reported evaluation metrics and indicate that ensemble-based machine learning can provide reliable early warning signals for sepsis using structured ICU data.

At the same time, the conclusions should be interpreted within the scope of the study design. The analysis was retrospective, based on a public benchmark dataset, and did not include external validation, fairness assessment, or direct comparison with modern temporal deep learning baselines. In addition, the feature engineering approach summarized time-series dynamics into fixed statistical descriptors, which may omit some fine-grained sequential information. Accordingly, the present work should be viewed as a reproducible and interpretable baseline framework that supports early sepsis prediction, while future studies should examine prospective validation, subgroup robustness, and comparison with newer temporal architectures.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Micheal O. Ajinaja, Yetunde E. Ogunwale , Johnson T. Fakoya , Akeem A. Abiona , Michael A. Ibiyomi , Oluyemisi A. Oyedemi

Acquisition, analysis, or interpretation of data: Micheal O. Ajinaja, Yetunde E. Ogunwale , Johnson T. Fakoya , Akeem A. Abiona , Michael A. Ibiyomi , Oluyemisi A. Oyedemi

Drafting of the manuscript: Micheal O. Ajinaja, Yetunde E. Ogunwale , Johnson T. Fakoya , Akeem A. Abiona , Michael A. Ibiyomi , Oluyemisi A. Oyedemi

Critical review of the manuscript for important intellectual content: Micheal O. Ajinaja, Yetunde E. Ogunwale , Johnson T. Fakoya , Akeem A. Abiona , Michael A. Ibiyomi , Oluyemisi A. Oyedemi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. All data generated or analyzed during this study are included in this published article and/or its appendices.

References

1. Rudd KE, Johnson SC, Agesa KM, et al.: Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet. 2020, 395:200-211. 10.1016/s0140-6736(19)32989-7

2. Singer M, Deutschman CS, Seymour CW, et al.: The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016, 315:801-810. 10.1001/jama.2016.0287

3. Matos Cunha LD, Ventura F, Pestana-Santos M, Mota M, Lomba L, Reis Santos M: Decision support strategies for bedside nursing clinical reasoning: A scoping review. International Journal of Nursing Studies Advances. 2025, 9:100393.

10.1016/j.ijnsa.2025.100393

4. Van Calster B, Collins GS, Vickers AJ, et al.: Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digital Health. 2025, 7:100916. 10.1016/j.landig.2025.100916

5. Tang Y, Zhang Y, Li J: A time series driven model for early sepsis prediction based on transformer module. BMC Medical Research Methodology. 2024, 24:23. 10.1186/s12874-023-02138-6

6. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K: Early prediction of sepsis in the ICU using machine learning: a systematic review. Frontiers in Medicine. 2021, 8:607952. 10.3389/fmed.2021.607952

7. Johnson AE, Bulgarelli L, Shen L, et al.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023, 10:1. 10.1038/s41597-022-01899-x

8. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H: TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature Communications. 2023, 14:7857. 10.1038/s41467-023-43715-z

9. Hassan R, Nguyen N, Finserås SR, Adde L, Strümke I, Støen R: Unlocking the black box: Enhancing human-AI collaboration in high-stakes healthcare scenarios through explainable AI. Technological Forecasting and Social Change. 2025, 219:124265. 10.1016/j.techfore.2025.124265

10. Abd-Alrazaq A, Solaiman B, Mekki YM, et al.: Hype vs reality in the integration of artificial intelligence in clinical workflows. JMIR Formative Research. 2025, 9:e70921. 10.2196/70921

11. Hur S, Lee Y, Park J, et al.: Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. npj Digital Medicine. 2025, 8:578. 10.1038/s41746-025-01958-8

12. Feretzakis G, Sakagianni A, Anastasiou A, et al.: Integrating Shapley values into machine learning techniques for enhanced predictions of hospital admissions. Applied Sciences. 2024, 14:5925. 10.3390/app14135925

13. Reyna MA, Josef CS, Jeter R, et al.: Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine. 2020, 48:210-217. 10.1097/ccm.0000000000004145

How to cite this article:
Ogunwale Y E, Fakoya J T, Oyedemi O A, et al. (March 24, 2026) Explainable Machine Learning for Multi-Horizon Early Sepsis Prediction in Intensive Care Units. Cureus J Comput Sci 3 : es44389-026-00045-7. DOI https://doi.org/10.7759/s44389-026-00045-7