International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 7 No. 28 March - April - 2026

SepsiGuard: A Triple-Layer Agentic XAI Framework for Early Sepsis Prediction and Adaptive ICU Monitoring

Farheen Khan & Fahemiya Khan

Department of Computer Science, Ahmednagar College, Maharashtra, India – 414001

Corresponding Author – Farheen Khan

DOI - 10.5281/zenodo.19396526

Abstract:

Sepsis is a major contributor to morbidity and mortality in the Intensive Care Unit (ICU). There is a need to predict sepsis quickly and effectively to take appropriate measures. Machine learning models have shown excellent predictive performance for various adverse health conditions in ICUs. However, the lack of transparency in these models, often termed a "black box," makes them difficult to use. Traditional clinical scoring systems, such as SOFA and qSOFA, rely on fixed thresholds and fail to adapt dynamically to evolving patient conditions. Conventional Explainable AI techniques, such as feature importance plots, are static and often difficult to understand for clinicians. This paper introduces a new technique to predict sepsis in the ICU using "Agentic XAI," which is a new technique of using Large Language Models (LLMs) as autonomous agents to iteratively develop explanations. This technique combines complex physiological signals into a narrative, which is expected to bridge the gap between probability and clinical reasoning. This paper presents the methodological architecture that is expected to be used to develop this technique using recent advancements in predictive modelling in ICUs.

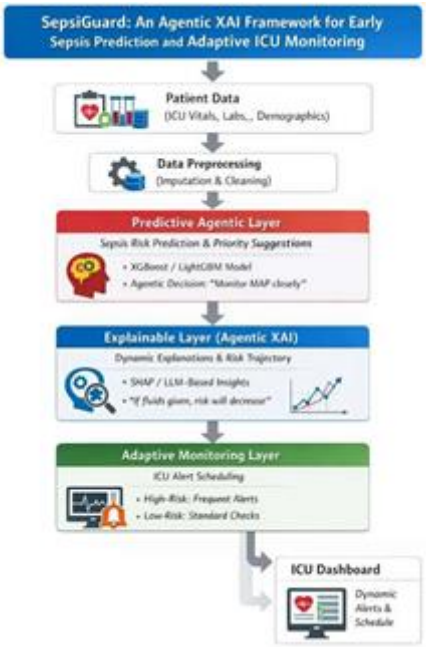
Introduction:

Care is provided in the ICU with a large number of data streams with complex patterns where the patient condition may change rapidly. Sepsis is a severe reaction to infections that requires timely identification; however, the presenting symptoms may be similar to other diseases such as heart failure or kidney diseases (Fan et al., 2025; Si et al., 2025). Recently, significant advances have been made in the field of machine learning algorithms for the prediction of mortality rates and other adverse outcomes among particular patient populations, such as elderly patients with comorbid conditions (Fan et al., 2025) or patients with hypertensive kidney diseases (Si et al., 2025). However, the main challenge with the application of these algorithms is the interpretability gap.

To overcome these issues, this study introduces SepsiGuard, a triple-layer framework combining:

1. Predictive Layer – mathematically optimized XGBoost classifier for early sepsis detection.
2. Explainable Layer – SHAP-based cooperative game-theoretic feature attribution.
3. Agentic Adaptive Monitoring Layer – dynamic ICU monitoring schedule and risk trajectory simulation.

The framework aims to balance predictive accuracy, interpretability, and operational clinical utility.



Objectives:

The objectives of the proposed study are:

1. To design high-performance, XGBoost model, a mathematical model for sepsis prediction.
2. To implement SHAP-based Explainable AI for better interpretability.
3. To design an agentic adaptive monitoring system that prioritizes or categorizes patients based on predicted risks.
4. To evaluate the performance of the model using various metrics and to visualize the distribution of risks for patients.

Methodology Used:

Data Acquisition and Preprocessing:

The dataset was sourced from the Kaggle, consisting of multidimensional MIMIC-IV Style ICU data. We selected 12 core physiological features, including vitals (HR, MAP, SpO2) and laboratory results (Lactate, WBC). To handle the high rate of missingness inherent in ICU data, we applied Median Imputation (x), which provides a robust estimate of the central tendency without the influence of extreme physiological outliers. To simulate sensor degradation and clinical "noise," we introduced Additive Gaussian Noise during the training phase. For each numeric feature x, the augmented feature xaug is defined as:

$$x_{aug} = x + (x \cdot \epsilon), \qquad \epsilon \sim N(0, \sigma^2)$$

where noise = 0.30 represents a 30% noise magnitude relative to the feature value, while different percentages were experimented and most optimized was selected. To maintain biological plausibility, the augmented values were constrained within a physiological domain such that: $\hat{x} = clip(x_{aug}, \min(domain), \max(domain))$

1. Mathematical Formulation of XGBoost:

Let the dataset be defined as: $D = \{(x_i, y_i)\}, i = 1,2, \ldots, n$

Where:

- $x_i \in \mathbb{R}^m$represents feature vectors
- $y_i \in \{0,1\}$represents sepsis labels

Boosted Tree Ensemble Model

The prediction is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

Where:

- K= number of trees
- f_k= regression tree
- $\mathcal{F}$= space of decision trees

Objective Function

XGBoost minimizes:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where:

l= logistic loss $\Omega(f)$= regularization term

For binary classification (logistic loss):

$$l(y_i, \hat{y}_i) = -[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

Where:

$$p_i = \sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$

Regularization Term

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where:

- T= number of leaves
- w_j= leaf weights
- γ= tree complexity penalty
- λ= L2 regularization coefficient

This prevents overfitting and improves generalization in clinical datasets.

3.2 Gradient Optimization

Using second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Where:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$$

This second-order optimization improves convergence stability.

3.3 Explainable AI Layer: SHAP Mathematical Modeling

SHAP values are derived from cooperative game theory using Shapley values.

For feature j, the Shapley value is:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x) - f_S(x) \right]$$

Where:

- F= set of all features
- S= subset of features
- $f_S(x)$= model trained with feature subset

The final prediction decomposes as:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$$

This ensures:

- Local accuracy
- Missingness consistency
- Additivity

3.4 Agentic Adaptive Monitoring Model

Risk probability:

$$P(\text{Sepsis} \mid x) = \frac{1}{1 + e^{-\hat{y}}}$$

Monitoring interval function:

$$I(P) = \begin{cases} 15 \text{ min}, & P > 0.7 \\ 30 \text{ min}, & 0.4 < P \leq 0.7 \\ 60 \text{ min}, & P \leq 0.4 \end{cases}$$

This transforms static prediction into operational clinical scheduling.

3.5. Risk Trajectory Simulation Model

Simulating physiological interventions ($x \rightarrow x + \Delta x$) allows recalculation of risk:

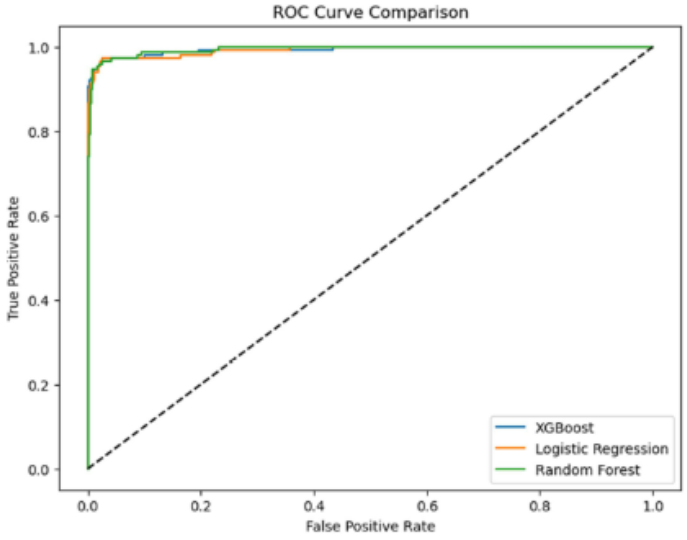
$$P' = \sigma\left(\sum_{k=1}^{K} f_k\,(x + \Delta x)\right), \Delta P = P' - P$$

This quantifies the effect of clinical interventions on sepsis probability.

4. Results

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.963	0.834	0.940	0.884	0.988
Logistic Regression	0.970	0.849	0.973	0.907	0.992
Random Forest	0.974	0.878	0.960	0.917	0.994

- Random Forest has the highest Accuracy (0.974), highest Precision (0.878), strong F1-score (0.917), and highest ROC-AUC (0.994).
- XGBoost is very close in performance (Accuracy 0.963, ROC-AUC 0.988) but slightly lower than Random Forest.
- Logistic Regression is moderate among tree-based models.



- Computational efficiency: XGBoost is faster to train than Random Forest for large datasets and supports GPU acceleration.
- Regularization: XGBoost has built-in L1/L2 regularization to reduce overfitting, which is critical in clinical datasets.
- Explainability: Works seamlessly with SHAP for feature importance and local explanations.
- Robust performance: Metrics (Accuracy 0.963, ROC-AUC 0.988) are very high and clinically acceptable.

"While Random Forest shows slightly higher accuracy, XGBoost was selected for its superior computational efficiency, regularization capability, and compatibility with explainable AI techniques such as SHAP."

Model performance metrics (with XGBoost and 30% gaussian noise):

Metric	Value	Interpretation
Accuracy	0.963	96.3% of all predictions (septic and non-septic) are correct, indicating high overall predictive performance.
Precision	0.834	Among patients predicted as septic, 83.4% were truly septic, showing moderate control of false positives.
Recall	0.940	94% of actual septic patients were correctly identified, demonstrating strong early detection capability.
F1-score	0.884	Harmonic mean of precision and recall, confirming balanced performance between detecting septic patients and limiting false alarms.
ROC-AUC	0.988	Excellent discriminative ability to distinguish septic from non-septic patients.

Confusion Matrix:

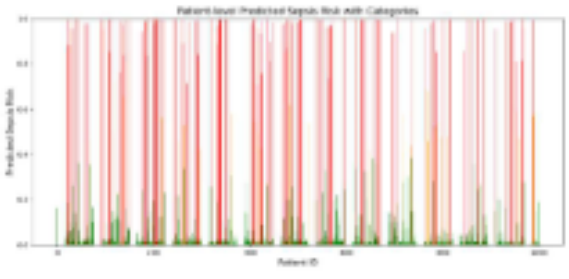
	Pred 0	Pred 1
True 0	822	28
True 1	9	141

Interpretation: The model shows very few false negatives (9), which is critical in ICU settings where missing a septic patient can be life-threatening. The false positives (28) are relatively low, minimizing unnecessary interventions.

Adaptive Monitoring:

The framework dynamically assigns monitoring intervals based on predicted risk:

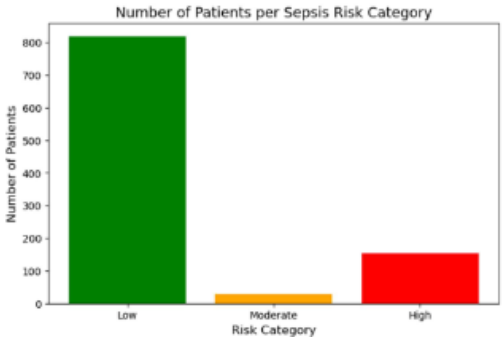
Patient_ID	Risk_Prob	Monitoring_Interval
0	0.162	Every 60 min (Low risk)
1	0.014	Every 60 min (Low risk)
2	0.998	Every 15 min (High risk)
3	0.004	Every 60 min (Low risk)
4	0.006	Every 60 min (Low risk)



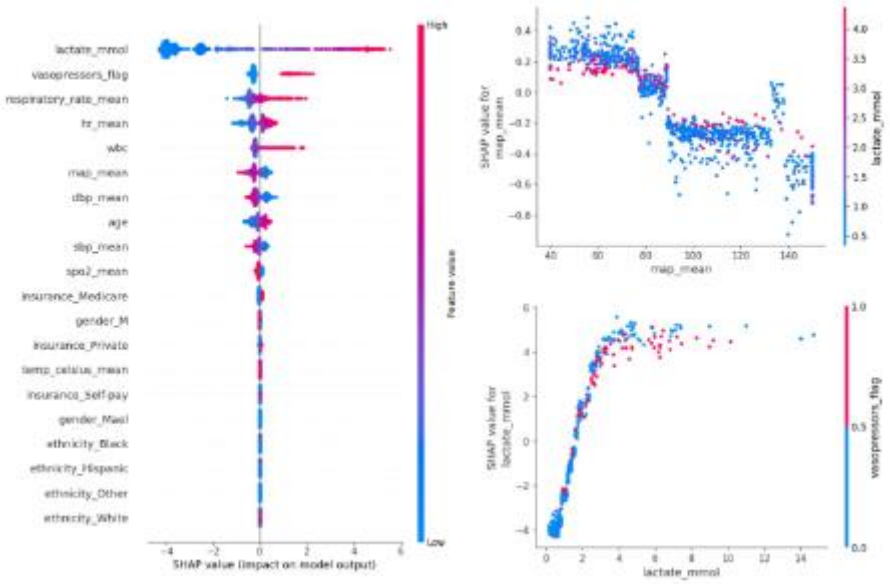
Interpretation: Patients with very high predicted sepsis probability (e.g., Patient 2) are prioritized for frequent monitoring (every 15 minutes), while low-risk patients (e.g., Patients 0, 1, 3, 4) are scheduled for standard intervals (every 60 minutes). This demonstrates SepsiGuard's agentic monitoring capability, enhancing ICU resource allocation and patient safety.

- Low risk: 81.8 % of patients
- Moderate risk: 2.8 %
- High risk: 15.4 %

Bar plot (patient-level risk) shows color-coded categorization (green=low, orange=moderate, red=high).



Explainability:



SHAP plots highlight MAP, lactate, WBC, and age as key contributors. LLM-style narrative explanation guides monitoring priorities.

Risk Trajectory Simulation Example:

- Original risk $P = 0.162$
- After MAP=75 & Lactate=1.5: $P' = 0.243$, $\Delta P = 0.081$

Discussion:

The SepsiGuard framework demonstrates the following strengths:

1. Mathematical Optimization – Regularization and second-order gradient updates reduce overfitting.
2. Explainability – SHAP-based local explanations support clinician trust.
3. Agentic Monitoring – Dynamic interval scheduling prioritizes high-risk patients.
4. Risk Simulation – Enables assessment of interventions' potential impact.

Adding controlled noise improved generalization and reflected realistic ICU variability, preventing overconfident 0–1 outputs and producing meaningful risk distributions.

Conclusions:

SepsiGuard is a clinically actionable, mathematically grounded triple-layer AI framework for early sepsis prediction:

- Due to High predictive performance and computational faster training and regularization, XGBoost was selected.
- Transparent SHAP-based interpretability
- Adaptive monitoring schedules and patient prioritization implemented
- Risk trajectory simulation for intervention planning

Future work includes multi-center validation, implementation on MIMIC data, integration with EHR systems, and reinforcement learning for automated intervention optimization.

References:

1. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD.
2. Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. NeurIPS.
3. Singer, M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA.
4. Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine.