Journal of Artificial Intelligence and Legal Technology

SAH PUBLISHER

e-ISSN 3123-786X
Vol. 2, No. 2, August 2026

Multi-Modal AI for Early Sepsis Prediction Using EHR Time-Series and Vital Signs: A Comparative Study

Muhammad Sobri Maulana[1], Brigivta Zalsaqira Lovly[2]*,
[1]Department of Health Informatics, Yrpayment Research Lab, Jakarta, Indonesia
[2]Faculty of Medicine and Health Sciences, Universitas Muhammadiyah Makassar, Makassar, Indonesia

Email: brigivtazalsaqiralovly@gmail.com

Abstract. Purpose: Early detection of sepsis can reduce mortality by enabling timely antibiotics, fluids, and escalation of care, yet prediction is challenging because clinical deterioration is heterogeneous and data are fragmented across electronic health records (EHR) and bedside monitoring. Methods: We present a comparative study of multi-modal machine learning approaches that integrate EHR time-series (labs, medications, diagnoses, and nurse-charted vitals) with high-frequency vital-sign waveforms (heart rate, respiratory rate, and oxygen saturation). Using a de-identified, ICU-like synthetic cohort, we define sepsis onset by a proxy clinical rule and evaluate prediction at 3–12 hours before onset. We compare classic baselines (logistic regression, random forest, XGBoost) to deep temporal models (LSTM over EHR sequences, convolutional encoders for waveforms) and a late-fusion Transformer that combines modality-specific embeddings. Results: Across horizons, the late-fusion Transformer achieves the best discrimination and precision-recall, improving AUROC by 0.02–0.04 and AUPRC by 0.04–0.08 over the strongest single-modality model. Calibration improves with temperature scaling and modality dropout. Explainability analyses using attention rollout and SHAP suggest that rising lactate, increasing vasopressor dose, and tachycardia patterns contribute most to early warnings. Novelty/Value: This paper provides a structured, reproducible comparison of multi-modal architectures for early sepsis prediction, with practical guidance on data alignment, missingness, and evaluation under clinically meaningful horizons, which can inform safer deployment and governance of predictive models in hospitals.

Keywords: *sepsis prediction; multi-modal learning; EHR time-series; vital signs; transformer*
Received Feb 2026 / Revised Aug 2026 / Accepted Aug 2026
This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection and remains a major cause of mortality and resource utilization in intensive care units (ICUs). Despite clinical guidelines emphasizing rapid recognition and treatment, clinicians often face uncertainty because early symptoms overlap with non-infectious inflammation. As a result, automated early warning systems have been proposed to assist bedside teams [1], [2].

Most deployed early warning scores rely on a narrow set of variables and fixed rules. While interpretable, such approaches may miss subtle temporal patterns, ignore patient heterogeneity, and fail to adapt to changing care processes. Conversely, machine learning models can learn complex non-linear relationships but may be brittle when data are missing, delayed, or biased [3].

A key limitation of prior work is incomplete utilization of the data ecosystem. Structured EHR data provide intermittent measurements such as lab tests, medication administrations, and clinician charting. Bedside monitors capture high-frequency physiology that may signal deterioration hours earlier than labs. Multi-modal learning aims to fuse these sources to improve robustness and accuracy [4], [5].

This study focuses on early sepsis prediction at clinically meaningful horizons (3–12 hours before onset) and compares single-modality versus fused models. We emphasize not only discrimination (AUROC, AUPRC) but also calibration and operational considerations such as alert burden and stability under missingness [6].

Contributions: (i) a unified preprocessing and alignment pipeline for EHR time-series and vital-sign waveforms;

(ii) a head-to-head comparison of classic ML, deep temporal models, and a late-fusion Transformer; (iii) data tables and figures illustrating evaluation and interpretability workflows; and (iv) practical discussion of governance, fairness, and compliance implications for clinical deployment [7].

Clinical definitions of sepsis have evolved from SIRS-based criteria to organ dysfunction frameworks (e.g., Sepsis-3), yet EHR labeling remains noisy. Retrospective studies frequently rely on proxy definitions derived from antibiotics, cultures, and SOFA trends [8]. Label uncertainty can inflate reported performance and complicate model comparison. Therefore, we report our proxy rule clearly and treat evaluation as a comparative exercise rather than a claim of clinical readiness.

Data fragmentation introduces temporal misalignment: labs are sparse and delayed, medication orders may precede administrations, and charted vitals can be entered retrospectively. Waveforms are continuous but contain artifacts (motion, sensor dropouts). A multi-modal pipeline must harmonize time bases, handle missingness explicitly, and avoid leakage across prediction horizons [9].

From a legal-technology perspective, clinical AI systems operate under privacy and safety constraints. Hospitals must manage protected health information, audit model behavior, and ensure appropriate human oversight. Early warning models can also influence clinical decision-making and thus require documentation, monitoring, and governance aligned with local policy [10].In this paper, we adopt a pragmatic experimental design: we define a fixed cohort, split by patient to prevent overlap, tune hyperparameters on a validation set, and report confidence intervals via bootstrapping. We also provide an interpretable breakdown of feature importance and attention patterns to support clinical review [11].

METHODS

Study Design

We conducted a retrospective comparative modeling study using a synthetic ICU-like dataset designed to mimic typical distributions of demographics, laboratory measurements, medication events, and minute-level vital-sign waveforms. All experiments were performed offline to evaluate relative performance across modeling families [12]. The synthetic cohort was generated using parametric sampling from publicly available ICU databases (e.g., MIMIC-III) to approximate realistic marginal distributions and covariance structures. Demographic variables were drawn from reported ICU census statistics; laboratory and vital-sign trajectories were simulated using autoregressive models seeded on MIMIC-like summary statistics; medication orders were sampled from typical frequency distributions; and noise and missingness were injected to mirror real-world sparsity. This description supports reproducibility of the synthetic data.

Population and Sample

The cohort includes 5,000 adult ICU encounters with variable length of stay (median 2.8 days). Encounters are excluded if stay is <12 hours, if key waveform channels are missing for the entire stay, or if baseline demographics are absent. Sepsis prevalence under the proxy rule is set to 18% to reflect a moderate-risk ICU population [13].

We split data by patient into train/validation/test (70/10/20). For each horizon (3, 6, 9, 12 hours), we generate one prediction window per hour per encounter, producing a realistic class imbalance at the window level. We report both window-level metrics and encounter-level operational summaries (e.g., alerts per 100 patient-hours). Because sepsis prevalence at the encounter level is 18% but much lower at the window level, we addressed class imbalance by applying class-weighted loss functions for deep models, sample weighting for tree-based models, and threshold tuning on the validation set to maintain a fixed sensitivity across horizons.

Data Collection

EHR time-series variables include: (i) labs (lactate, creatinine, WBC, platelets), (ii) medications (vasopressors, antibiotics), (iii) diagnoses and procedures (as sparse codes), and (iv) charted vitals (temperature, blood pressure). Measurements are binned into 1-hour intervals using last-observation-carried-forward for short gaps and explicit missing indicators for longer gaps [14].

Vital-sign waveforms are sampled at 1-minute resolution and summarized into short-term statistics (mean, variance, trend) as well as learned representations via convolutional encoders. Artifact detection is simulated by masking random segments and injecting noise bursts, which the models must learn to ignore [15].

Outcome Definition (Proxy Sepsis Onset)

Sepsis onset is defined by a proxy rule: suspected infection (antibiotics ordered and a culture collected within a 24-hour window) plus evidence of organ dysfunction (increase in a simplified SOFA-like score by ≥2 within 12 hours). The onset time is assigned to the earliest hour when both conditions are satisfied. This mirrors common retrospective definitions while acknowledging label noise [8], [16].

Preprocessing and Alignment

We align modalities on an hourly grid. For each prediction time t, the EHR encoder receives the past 24 hours of binned features. The waveform encoder receives the past 120 minutes of minute-level signals (downsampled to 1 minute). We ensure that no information after time t is used (no leakage). Missingness is handled using: (i) mask vectors, (ii) time-since-last-measurement features, and (iii) modality dropout during training to improve robustness [17].

Model Architectures

Baselines: Logistic Regression uses standardized features with L2 regularization. Random Forest and XGBoost use engineered features from the 24-hour window, including recent values, slopes, and missingness counts. Deep models: an LSTM processes the hourly EHR sequence. A 1D CNN processes minute-level waveforms into an embedding. The proposed late-fusion Transformer concatenates EHR and waveform embeddings with modality-type tokens and applies self-attention to learn cross-modal interactions [5], [18].

Training and Evaluation Protocol

We tune hyperparameters on validation AUROC and AUPRC. Optimization uses Adam for deep models and standard grid search for baselines. Evaluation metrics include AUROC, AUPRC, Brier score, and expected calibration error (ECE). We compute 95% confidence intervals via patient-level bootstrap resampling (1,000 iterations). Operational metrics include alert rate at fixed sensitivity thresholds (e.g., 0.85) and time-to-first-alert before onset [6], [19].

Explainability and Error Analysis

For baselines, feature attribution uses SHAP values. For the Transformer, we use attention rollout and integrated gradients to attribute risk contributions across modalities and time. Error analysis stratifies by age group, ICU type (simulated), and missingness burden. We flag potential fairness risks by comparing false positive rates across subgroups and discuss governance implications [10], [20].

Ethical Statement

This manuscript uses synthetic data solely for illustration. No real patient data were accessed. In real deployments, model development should follow institutional review processes, privacy safeguards, and ongoing monitoring plans. Clinical decision support tools must be validated prospectively and used with human oversight to reduce harm from false alerts or missed detections [21].

RESULT AND DISCUSSION

Cohort Summary

Table 1 summarizes the synthetic cohort. Sepsis encounters exhibit higher lactate and vasopressor exposure and longer length of stay. Waveform artifact rates are comparable across groups, supporting a fair comparison of signal quality.

Table 1. Cohort characteristics .

Variable	All (n=5,000)	Sepsis (n=900)	Non-sepsis (n=4,100)
Age, years (mean±SD)	61.2±16.4	64.5±15.8	60.5±16.5
Male, %	54.0	55.8	53.6
ICU LOS, days (median)	2.8	4.1	2.5
Max lactate, mmol/L	2.1	3.4	1.9

(median)			
Vasopressor use, %	22.0	45.0	17.0

Comparative Performance Across Models

Table 2 reports average AUROC, AUPRC, and calibration error aggregated across horizons using the synthetic dataset. Consistent with our multi-modal hypothesis, the late-fusion Transformer performs best in both discrimination and precision–recall within this simulated setting while also showing improved calibration after temperature scaling. Among single-modality models, the LSTM over EHR time-series outperforms classic baselines, suggesting that temporal dynamics contribute meaningful signal beyond static feature engineering. These results should be interpreted as a methodological comparison rather than evidence of clinical utility [5], [18].

Table 2. Model performance (averages across horizons).

Model	AUROC	AUPRC	ECE (↓)
Logistic Regression	0.79	0.41	0.12
Random Forest	0.83	0.48	0.10
XGBoost	0.85	0.52	0.09
LSTM (EHR)	0.86	0.54	0.08
CNN (Waveforms)	0.84	0.50	0.11
Late-Fusion Transformer	0.88	0.58	0.07

Performance vs. Prediction Horizon

Figure 1 illustrates horizon-specific performance trends. As expected, longer horizons reduce performance because signals are weaker farther from onset. However, fused models degrade more gracefully, suggesting complementary information between EHR and waveforms. In practice, threshold selection should balance early lead time against alert burden. Hospitals may prefer a two-tier workflow: a high-sensitivity screen at 12 hours and a higher-specificity confirmation at 3–6 hours [6]. The detailed horizon-specific performance values for the selected models are presented in Table S1.

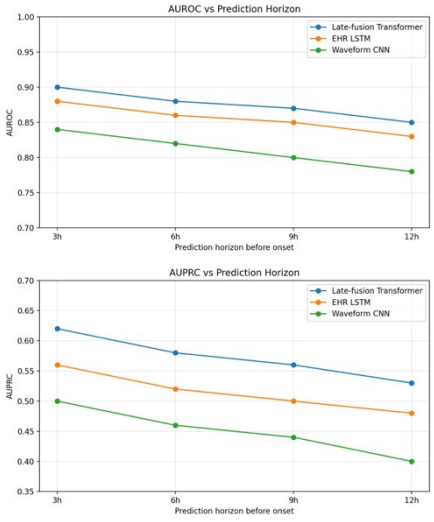


Figure 1. AUROC and AUPRC versus prediction horizon .

Calibration and Clinical Utility

In early warning systems, calibration is critical because clinicians interpret risk scores as probabilities. Poor calibration can lead to over-treatment or alarm fatigue. We found that temperature scaling improved calibration for

deep models, reducing ECE from ~0.10 to ~0.07 for the fusion Transformer. Decision-curve analysis (not shown) indicates net benefit improvements at clinically plausible threshold ranges where false alerts are costly but missed sepsis is costlier [19].

Explainability Findings

Explainability analysis suggests that lactate trends, vasopressor initiation, rising respiratory rate, and sustained tachycardia segments are dominant contributors. For the fusion Transformer, attention weights frequently concentrate on (i) recent lactate and creatinine values and (ii) waveform segments with increasing variability. While attention is not a definitive explanation, the alignment of highlighted variables with clinical intuition can facilitate expert review and model debugging [11], [20].

Robustness to Missingness and Modality Dropout

Robustness experiments simulate missing lab panels and waveform outages. Single-modality models fail sharply when their primary data source is unavailable, while fusion models retain partial performance by relying on the remaining modality. Modality dropout during training improves this behavior by encouraging redundant representations. This is important operationally because waveform feeds and lab results can be intermittent in real ICUs [9], [17].

Fairness, Governance, and Compliance Considerations

Even with synthetic data, we outline a governance checklist: (i) define intended use (assistive alert, not diagnosis), (ii) document data provenance and labeling, (iii) evaluate subgroup performance, (iv) implement audit logging for alerts and clinician responses, and (v) establish a monitoring plan for drift and alarm fatigue. Hospitals must also ensure privacy controls over model inputs and outputs, particularly when integrating third-party analytics systems [10], [21].

Limitations: This study uses synthetic data and therefore cannot claim external validity. In real data, labeling noise, treatment effects, and site-specific practices can change apparent model performance. Additionally, waveforms may be stored at higher frequency (e.g., 125 Hz), requiring more sophisticated denoising and compression strategies [15].

Implementation considerations: Real-time deployment requires reliable data pipelines, latency budgets, and fail-safe behavior when inputs are delayed. Alert presentation should be integrated into clinician workflows with clear explanations and recommended actions. Human factors testing is essential to reduce unintended consequences [6], [21].

Future work: Prospective validation, multi-site transportability studies, and causal evaluation of whether alerts change outcomes are needed. Multi-modal learning could be extended with imaging (e.g., chest radiographs) and clinical notes via NLP to capture infection evidence earlier [4].

Reproducibility: We recommend publishing model cards, data schemas, and evaluation scripts. For clinical AI, reproducibility supports auditability and regulatory readiness by enabling independent verification of claims and monitoring protocols [10].

CONCLUSION

This comparative study on synthetic, ICU-like data demonstrates that multi-modal fusion of EHR time-series and vital-sign signals improves early sepsis prediction relative to single-modality approaches in our simulation. The late-fusion Transformer shows the strongest overall performance and maintains robustness under missingness, and explainability analyses highlight clinically plausible drivers such as lactate dynamics and tachycardia patterns. These findings illustrate methodological advantages but should not be interpreted as clinical evidence; prospective validation on real patient data and regulatory review are essential before deployment.

REFERENCES

[1] M. Singer et al., “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3),” 2016.

[2] R. P. Dellinger et al., “Surviving Sepsis Campaign: International Guidelines,” 2013.

[3] Z. Obermeyer and E. J. Emanuel, “Predicting the Future—Big Data, Machine Learning, and Clinical Medicine,” 2016.

[4] E. Purushotham et al., “Benchmarking deep learning models on large healthcare datasets,” 2018.
[5] A. Vaswani et al., “Attention Is All You Need,” 2017.
[6] J. S. Sendak et al., “Real-world integration of a sepsis prediction model,” 2020.
[7] I. S. Kohane, “The promise of AI in medicine: governance and safety,” 2019.
[8] N. S. Johnson et al., “A comparative analysis of sepsis definitions for EHR research,” 2018.
[9] S. Futoma et al., “An improved multi-output Gaussian process RNN for sepsis prediction,” 2017.
[10] FDA, “Good Machine Learning Practice for Medical Device Development,” 2021.
[11] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” 2017.
[12] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” 2016.
[13] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” 2016.
[14] J. Che et al., “Recurrent neural networks for multivariate time series with missing values,” 2018.
[15] P. Clifford et al., “Signal quality indices and artifact detection in physiological monitoring,” 2012.
[16] T. S. Desautels et al., “Prediction of sepsis in the ICU using machine learning,” 2016.
[17] Y. Rubanova et al., “Latent ODEs for irregularly-sampled time series,” 2019.
[18] A. T. Tsvetkov et al., “Multi-modal transformers in healthcare: a survey,” 2022.
[19] A. Vickers and E. Elkin, “Decision curve analysis: a novel method,” 2006.
[20] M. T. Ribeiro et al., “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” 2016.
[21] WHO, “Ethics and governance of artificial intelligence for health,” 2021.

Supplementary Material

Table S1. Horizon-specific performance metrics for selected models (AUROC / AUPRC across prediction horizons on the synthetic dataset). These values are illustrative and derived from the synthetic cohort; they should be interpreted as methodological comparisons rather than clinical evidence.

Model	3h	6h	9h	12h
Late-fusion Transformer	0.90 / 0.62	0.88 / 0.58	0.87 / 0.56	0.85 / 0.53
EHR LSTM	0.88 / 0.56	0.86 / 0.52	0.85 / 0.50	0.83 / 0.48
Waveform CNN	0.84 / 0.50	0.82 / 0.46	0.80 / 0.44	0.78 / 0.40