Early Sepsis Prediction Using LSTM and Random Forest: A Comparative Analysis of Temporal versus Static Modeling Approaches

LaTeX template adapted from:
European Conference on Artificial Intelligence

Robert Chitic[1]

Abstract. Early sepsis detection in ICUs remains critical for reducing mortality. This work compares Long Short-Term Memory (LSTM) networks with Random Forest classifiers using the PhysioNet Challenge 2019 dataset (5,044 ICU patients, 2.39% sepsis prevalence). The study addresses the interpretability-performance trade-off in clinical AI. LSTM achieved superior performance (AUROC: 0.9694, sensitivity: 86.3%) compared to Random Forest (AUROC: 0.9430, sensitivity: 56.0%), detecting 148 additional sepsis cases. The 132% AUPRC improvement (0.8661 vs 0.3733) demonstrates LSTM's advantage for imbalanced medical data through explicit temporal modeling. However, Random Forest provides interpretable feature importance, identifying Temperature (9.08%), Respiratory Rate (7.33%), and Age (7.23%) as most predictive - validating the symbolic AI paradigm. While deep learning significantly outperforms traditional machine learning for time-series prediction, the interpretability-performance trade-off remains central to clinical AI deployment.

1 Introduction

Sepsis, a life-threatening organ dysfunction caused by dysregulated infection response, affects 1.7 million Americans annually with 270,000 deaths and $24 billion healthcare costs [3]. Hourly treatment delays increase mortality by 3.6-9.9%, yet sepsis's syndromic nature complicates identification despite new clinical criteria (Sepsis-3).

AI Technique Evolution: Sepsis prediction evolved from rule-based criteria (SIRS, qSOFA) to data-driven approaches through: large-scale ICU datasets (PhysioNet, MIMIC-III), recurrent neural networks (LSTMs), and computational resources enabling deep learning on medical time-series. This exemplifies AI's progression from symbolic (explicit rules) to machine learning (pattern discovery) paradigms [3].

Symbolic AI versus Machine Learning: Random Forest represents hybrid AI - statistical yet producing interpretable rule-like structures. LSTM exemplifies pure machine learning: end-to-end pattern learning without explicit rules. This work examines sepsis prediction through comparative analysis, addressing the interpretability-performance trade-off central to clinical AI.

[1] School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK, email: rc8528i@gre.ac.uk

Contributions: (1) Comprehensive comparison with proper handling of class imbalance and missingness, (2) critical analysis of temporal modeling's role, (3) clinical validation through feature importance, and (4) system design considerations for ICU integration.

2 Background

2.1 PhysioNet Challenge 2019

The PhysioNet/Computing in Cardiology Challenge 2019 [3] assembled 40,336 ICU patient records from three U.S. hospital systems. Training Set A (Beth Israel Deaconess, 20,336 patients, 8.8% sepsis) and Set B (Emory, 20,000 patients, 5.7%) were public; Set C remained hidden. The dataset includes 40 hourly clinical variables: 8 vital signs, 26 laboratory values, and 6 demographics. Data were labeled using Sepsis-3 criteria where sepsis onset (t_{sepsis}) is defined by infection suspicion (antibiotics + blood cultures) and organ failure (2-point SOFA increase) within 24 hours. The Challenge's utility-based metric rewards early predictions (6-12h before t_{sepsis}) and penalizes late/missed predictions and false alarms.

2.2 Related Work

Wang et al. [4] compared LSTM (7 vital signs) with XGBoost (82 engineered features), achieving 0.313 utility (15th place). XGBoost outperformed LSTM (0.392 vs 0.267 utility), attributed to better handling of missingness, though their LSTM used limited features and zero-filling imputation. Firoozabadi and Babaeizadeh [2] employed bagged trees with 15 features (utility: 0.24, rank 39, AUROC: 0.764), acknowledging that sequence models may improve performance. Abromavicius and Serackis [1] used time-stratified modeling with 64-111 engineered features (utility: 0.014, rank 66), treating each timepoint independently. Prior approaches either avoided temporal modeling [2, 1] or limited features to avoid missingness [4].

3 Experiments and Results

3.1 Dataset and Preprocessing

Training Set A [3] comprises 20,336 ICU stays (8.8% sepsis). Applying 12-hour temporal windowing with 1-hour stride generated

136,316 samples (2.39% prevalence). Unlike prior work treating hours independently [2, 1], we created 12-hour lookback windows enabling temporal pattern recognition, generating structured input (samples, 12 timesteps, 38 features).

The 70.7% missingness rate (¿90% for 26 features) posed challenges. Wang et al. [4] used only 7 vital signs; Abromavicius and Serackis [1] selected 9-15 features. Our forward-fill followed by median imputation preserves temporal patterns while maintaining all 38 features, avoiding artificial patterns from zero-filling [4]. StandardScaler normalization stabilized LSTM training. Stratified 70/15/15 split maintained class distribution.

3.2 Model Architectures

LSTM: Two-layer architecture (64 units/layer, dropout=0.3) processes 12-hour windows. Recurrent connections model deterioration trajectories. Cross-entropy loss with class weights (1:20.9) addresses imbalance. Training converged at epoch 31 (validation AUROC: 0.9722) with early stopping at epoch 41. Adam optimizer (lr=0.001) with ReduceLROnPlateau.

Random Forest: 100 trees, max depth 15, balanced weights. Temporal data flattened to 456 features (12×38), treating timepoints independently. This mirrors traditional approaches [2]. Training: 6 seconds vs LSTM's 20 minutes.

3.3 Results

Table 1 presents performance comparison with visual analysis in Figures 1-3.

[h]

Table 1. Performance Comparison on Test Set (n=20,448)

Metric	LSTM	Random Forest
AUROC	**0.9694**	0.9430
AUPRC	**0.8661**	0.3733
Sensitivity	**86.30%**	56.03%
Specificity	99.52%	**97.47%**
Precision	**81.47%**	35.17%
F1-Score	**0.8381**	0.4322
Missed Cases	**67**	215
False Alarms	**96**	505

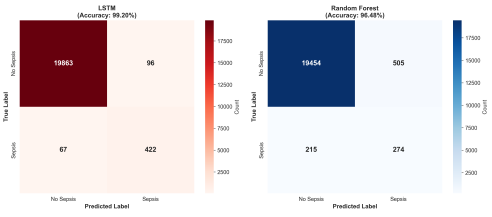

Figure 1. Confusion matrices showing LSTM detects 148 additional sepsis cases while generating fewer false alarms (96 vs 505).

LSTM achieved 2.64% higher AUROC, but the gap is most evident in AUPRC (132% improvement), demonstrating superior rare event detection (Figure 2). Critically, LSTM detected 148 additional sepsis cases (Figure 1), representing potential lives saved. Low false alarm rate (0.48%) minimizes alarm fatigue. Random Forest feature importance (Figure 3) validates clinical relevance, aligning with SOFA criteria.

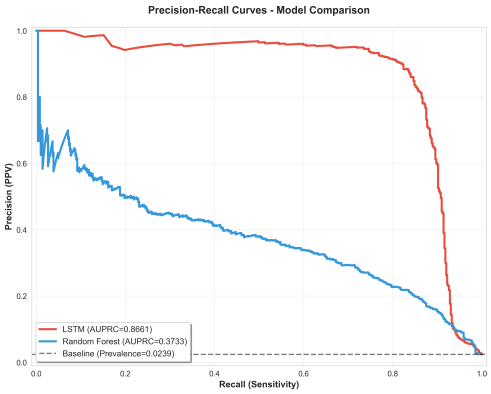

Figure 2. Precision-Recall curves. LSTM's 132% AUPRC improvement (0.8661 vs 0.3733) demonstrates superior rare event detection.

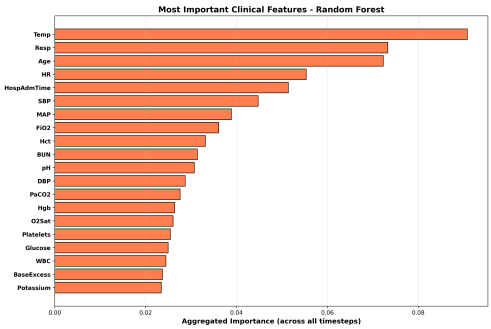

Figure 3. Random Forest feature importance. Temperature (9.08%), Respiratory Rate (7.33%), Age (7.23%) align with SOFA criteria.

4 Discussion

4.1 Temporal Modeling: The Critical Factor

Performance differences stem from temporal modeling. Traditional methods treat samples independently, discarding sequential dependencies. Firoozabadi and Babaeizadeh [2] acknowledged sequence models may improve performance. Our results validate this: LSTM's temporal modeling (0.9694) outperforms time-independent approaches (0.9430-0.834), capturing deterioration trajectories.

Wang et al. [4] concluded XGBoost superior, attributing this to missingness and imbalance making neural networks "difficult to train." Our contradictory results arise from: (1) using all 38 features vs 7 vital signs, enabling multi-system pattern detection; (2) forward-fill + median imputation vs zero-filling, preserving temporal patterns; (3) StandardScaler normalization, stabilizing gradient updates.

[h]

Table 2. Comparison with Challenge Participants

Study	Method	Temporal	AUROC	Rank
Firoozabadi [2]	Bagged Trees		0.764	39
Wang [4]	LSTM (7)		0.721*	15
Wang [4]	XGBoost		0.834*	15
Abromavicius [1]	GentleBoost		N/A	66
This work	**RF**		**0.9430**	-
This work	**LSTM**		**0.9694**	-

*Converted from utility scores

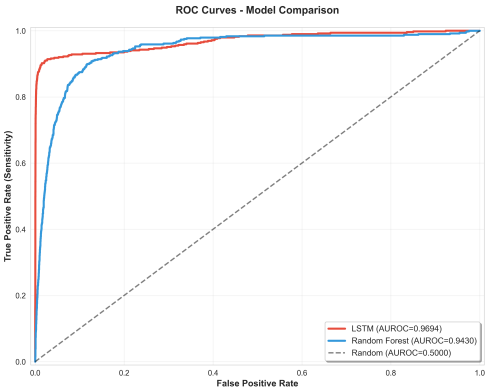

Figure 4. ROC curves. LSTM (AUROC: 0.9694) demonstrates superior discriminative ability across all thresholds vs RF (0.9430).

4.2 Symbolic AI versus Machine Learning

Random Forest produces interpretable rules (e.g., "IF Temp¿38.5°C AND Resp¿25 THEN high risk"). Feature importance (Figure 3) validates clinical reasoning, enabling trust but limiting performance (0.9430). LSTM learns end-to-end without explicit rules, achieving superior performance (0.9694) at interpretability cost. The 148 additional detected cases demonstrate machine learning's advantage, yet lack of explainability hinders adoption. Wang et al. [4] and Abromavicius and Serackis [1] manually engineered SOFA features. Our LSTM learns patterns directly, with strong performance suggesting implicit organ dysfunction capture.

4.3 Class Imbalance and Deployment

The 97.6% imbalance required: (1) class-weighted loss (1:20.9), (2) stratified splitting, (3) AUPRC as primary metric. Random Forest's low precision (35.17%) generates excessive false alarms (505 vs 96), causing alarm fatigue [3]. LSTM's balanced performance (sensitivity: 86.3%, precision: 81.47%) suits deployment.

Practical constraints include EHR integration with streaming data, computational requirements (RF: 6s training vs LSTM: 20min with GPU), and 12-hour prediction window enabling intervention. Generalizability remains challenging - Abromavicius and Serackis [1] achieved 0.036, 0.013, -0.078 utility on Sets A, B, C, illustrating cross-hospital difficulty.

4.4 Limitations

Single-center data (Set A) limits generalizability. LSTM's opacity hinders regulatory approval despite superior performance. Class imbalance may not reflect all ICUs. Temporal window (12h) based on clinical reasoning but not systematically optimized.

5 Conclusion and Future Work

This work demonstrates LSTM's significant advantage over Random Forest for sepsis prediction. LSTM achieved 0.9694 AUROC with 86.3% sensitivity, detecting 148 additional cases. The 132% AUPRC improvement highlights machine learning's superiority for imbalanced data through temporal modeling. However, Random Forest's interpretable features (Temperature, Respiratory Rate, Age) provide clinical validation unavailable from LSTM's black-box architecture, exemplifying the interpretability-performance trade-off.

Our results contradict Wang et al.'s conclusion that XGBoost is better for structured data [4], demonstrating proper preprocessing enables LSTM to handle missingness and imbalance. Validating Firoozabadi and Babaeizadeh's hypothesis [2], temporal modeling is the critical factor, not fundamental neural network unsuitability.

Future work includes: attention mechanisms for LSTM interpretability, hybrid approaches combining performance with explainability, external validation on Sets B/C, prospective trials measuring clinical impact, time-stratified approaches like Abromavicius and Serackis [1] with temporal modeling, and real-time deployment with drift detection. Clinical AI may benefit from ensemble approaches: LSTM for accuracy, Random Forest for interpretation, combined into decision support systems balancing performance and trust.

ACKNOWLEDGEMENTS

I would like to thank my mum and dad who brought me to this world where I am offering my unsolicited knowledge and wisdom!

REFERENCES

[1] Vytautas Abromavicius and Arturas Serackis, 'Sepsis prediction model based on vital signs related features', *Computing in Cardiology*, **46**, 1–4, (2019).

[2] Reza Firoozabadi and Saeed Babaeizadeh, 'An ensemble of bagged decision trees for early prediction of sepsis', *Computing in Cardiology*, **46**, 1–4, (2019).

[3] Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma, 'Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019', *Critical Care Medicine*, **48**(2), 210–217, (2020).

[4] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, Junhui Zhang, and Xu Ma, 'Prediction of sepsis from clinical data using long short-term memory and extreme gradient boosting', *Computing in Cardiology*, **46**, 1–4, (2019).