Early Sepsis Detection with Machine Learning

Insights from MIMIC-III Data

Research Report by
Jesus Minjares
The University of Texas at Austin
jesusminjaresjr@utexas.edu
Date: April 22, 2025

Additional Resources

Resource	Link
Presentation and Code	https://github.com/jminjares4/AI-in-Healthcare/tree/main/High_Risk_Project
Video Presentation	https://utexas.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2336ffb3-4ae8-47af-bb7f-b2c4016a75ec

Early Sepsis Detection with Machine Learning: Insights from MIMIC-III Data

Jesus Minjares, *The University of Texas at Austin*

***Abstract*—Sepsis, a life-threatening condition with a mortality rate of approximately 30%, underscores the importance of early identification to enhance patient outcomes. This study developed a machine learning framework that integrates clinical notes and vital signs from the MIMIC-III database. Text data from medical notes were processed using BioBERT, and two models, LSTM and Logistic Regression, were trained, achieving an AUROC of 0.737. Significant predictors, such as heart rate and references to infection, were identified; however, further refinement is required to improve predictive accuracy and ensure equitable performance across diverse patient groups. Recent research highlights the potential of machine learning in sepsis prediction, yet challenges persist due to inconsistent clinical data. Future efforts will prioritize real-time monitoring and equitable model performance across varied populations.**

***Index Terms*—Sepsis detection, machine learning, BioBERT, LSTM, fairness evaluation**

I. Introduction

Sepsis arises from an excessive immune response to infection, resulting in organ dysfunction and a substantial risk of mortality. In the United States, it affects 1.7 million adults annually, contributing to 270,000 deaths and incurring healthcare costs exceeding $62 billion [1], [2]. Timely detection is critical, as each hour of delayed treatment increases mortality risk by 7.6% [3]. Conventional approaches, such as qSOFA and SIRS, fail to identify up to 30% of cases due to limited sensitivity [4]. To address this challenge, a machine learning framework was developed using the MIMIC-III dataset, integrating vital signs and clinical notes. By employing BioBERT and LSTM models, this study seeks to enhance prediction accuracy and promote fairness across patient demographics, facilitating prompt interventions and reducing mortality.

II. Related Work

Advancements in machine learning for sepsis prediction have informed this research. Yadgarov et al. (2024) reviewed 36 studies, observing that neural networks and decision trees frequently achieved AUROC scores above 0.80, surpassing traditional metrics like SOFA [5]. They highlighted inconsistent data quality as a persistent challenge, a difficulty also encountered in this study. Hu et al. (2024) evaluated COMPOSER, a deep-learning model implemented in emergency departments, which enhanced treatment protocols and reduced in-hospital mortality through a Best Practice Advisory system, though further validation is required [6]. Zhou et al. (2025) proposed an interpretable model for sepsis risk assessment at triage using MIMIC-IV, attaining an AUROC of 0.83 with Gradient Boosting and emphasizing detailed triage data [4]. This study distinguishes itself by incorporating clinical notes and prioritizing fairness, addressing disparities in equitable care and real-time applicability.

III. Methodology

The MIMIC-III dataset, encompassing vital signs (CHARTEVENTS), clinical notes (NOTEEVENTS), diagnoses (DIAGNOSES_ICD), and patient demographics (PATIENTS), served as the data source. The methodology comprised four phases:

1) **Data Acquisition**: Vital signs, clinical notes, diagnoses, and patient information were collected from MIMIC-III.
2) **Data Preprocessing**: Vital signs, including heart rate, blood pressure, and respiratory rate, were aggregated into 12-hour intervals. Missing values were imputed using the last observed value or the dataset mean. Clinical notes were analyzed for terms such as "infection" and converted into BioBERT embeddings. Sepsis labels were derived from ICD9 codes.
3) **Model Development**: Logistic Regression and LSTM models were trained over five epochs. The LSTM architecture included 64 units, a 32-unit dense layer, and a sigmoid activation function.
4) **Performance Evaluation**: Models were assessed using AUROC, precision, recall, and F1 scores. Fairness across male and female patients was examined, and SHAP values identified the most influential features.

IV. Results

The LSTM model achieved an AUROC of 0.737, marginally outperforming Logistic Regression's 0.722, as depicted in the ROC curves in Figure 1. Confusion matrices in Figure 2 indicate that the LSTM accurately classified 24,233 non-sepsis cases but failed to detect 16,432 sepsis cases, whereas Logistic Regression missed fewer sepsis cases (10,981) but generated 8,335 false positives. Precision-recall curves in Figure 3 reveal that Logistic Regression exhibited higher precision at lower recall, while LSTM performed better at higher recall. Fairness analysis in Figure 4 shows comparable AUROC scores (0.731 for females, 0.743 for males), yet recall was higher for females (0.197) than males (0.172), suggesting potential bias. SHAP analysis in Figure 5 identified heart rate (SHAP values up to 0.2), infection mentions (up to 0.8), and fever (up to 0.1) as primary predictors. Vital sign distributions in Figure 6 demonstrate that sepsis patients exhibited elevated heart rates (100–120 bpm versus 80–100 for non-sepsis), lower blood

pressure (occasionally under 80 mmHg versus 80–100), and higher respiratory rates (20–25 breaths/min versus 15–20). BioBERT embeddings are visualized in Figure 7, with t-SNE analysis in Figure 8 revealing clusters (e.g., Notes 2 and 23) reflecting similar clinical note patterns. Training metrics in Figure 9 indicate that the LSTM attained a training accuracy of 0.7505 and validation accuracy of 0.7495, with training and validation losses of 0.500 and 0.502, respectively. These findings underscore the framework's potential for early sepsis detection, though further improvements in accuracy and fairness are necessary.

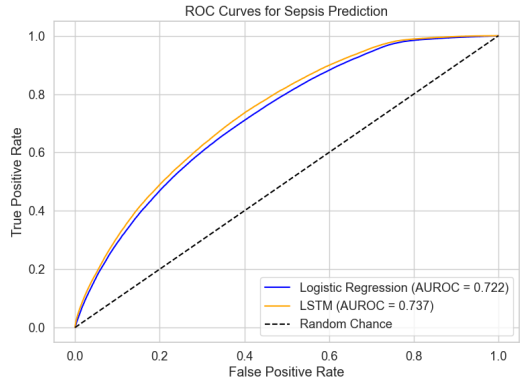

Fig. 1. ROC Curves for Sepsis Prediction (LSTM: AUROC 0.737, Logistic Regression: AUROC 0.722).

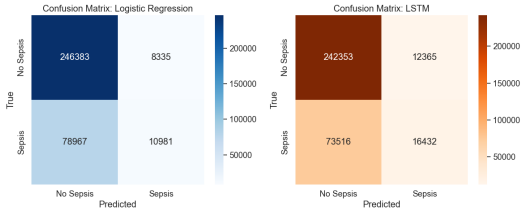

Fig. 2. Confusion Matrices for Sepsis Prediction. Logistic Regression: 24,683 true negatives, 10,981 false negatives; LSTM: 24,233 true negatives, 16,432 false negatives.

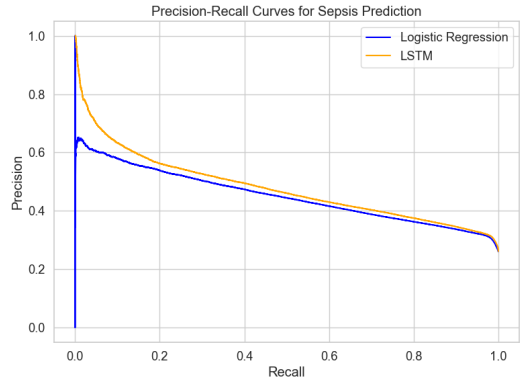

Fig. 3. Precision-Recall Curves for Sepsis Prediction.

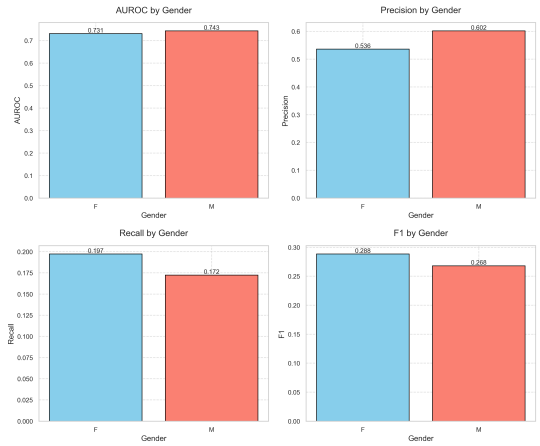

Fig. 4. Fairness Metrics by Gender. AUROC: Female 0.731, Male 0.743; Recall: Female 0.197, Male 0.172.

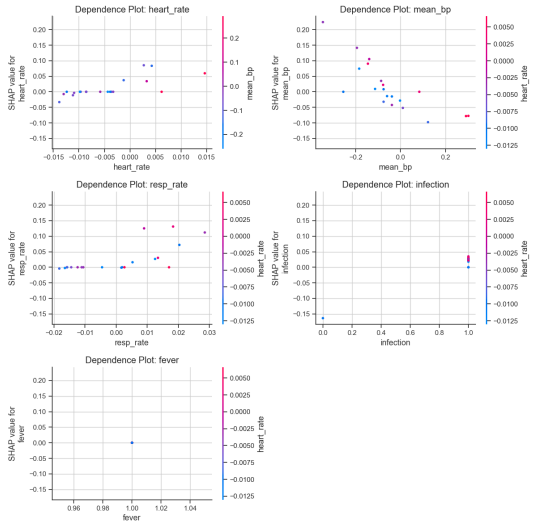

Fig. 5. SHAP Dependence Plots. Heart Rate: SHAP values up to 0.2; Infection: SHAP values up to 0.8; Fever: SHAP values up to 0.1.

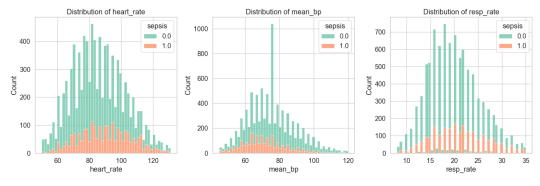

Fig. 6. Vital Sign Patterns. Heart Rate: Sepsis 100–120 bpm, Non-Sepsis 80–100 bpm; Blood Pressure: Sepsis some ¡80 mmHg, Non-Sepsis 80–100 mmHg; Breathing Rate: Sepsis 20–25 breaths/min, Non-Sepsis 15–20 breaths/min.

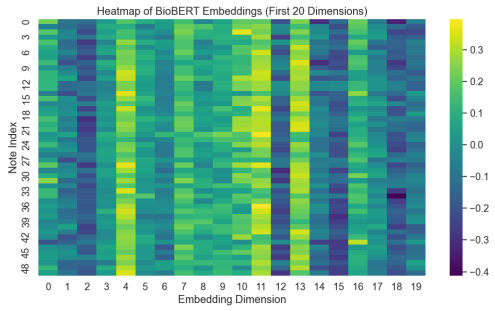

Fig. 7. Heatmap of BioBERT Embeddings (First 20 Dimensions). Values range from -0.4 to 0.3, reflecting note semantics.

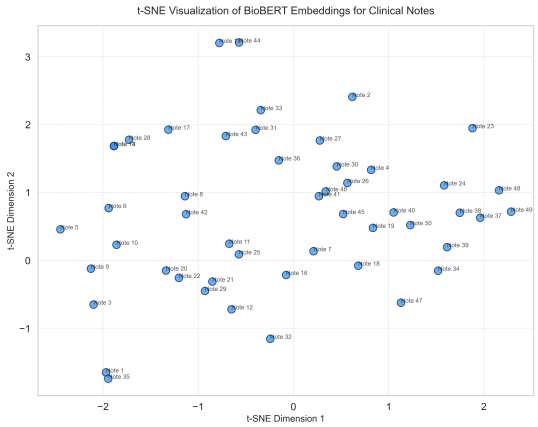

Fig. 8. t-SNE Visualization of BioBERT Embeddings for Clinical Notes.

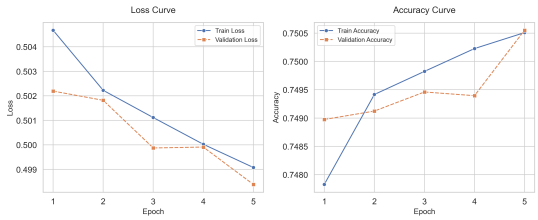

Fig. 9. LSTM Training Results. Accuracy: Training 0.7505, Validation 0.7495; Loss: Training 0.500, Validation 0.502.

V. Conclusion

A machine learning framework for early sepsis detection was developed, achieving an AUROC of 0.737 using LSTM and identifying heart rate and infection mentions as critical predictors. The integration of BioBERT embeddings with vital signs enhanced predictive performance, yet improvements in sensitivity and fairness are required. The t-SNE visualization in Figure 8 suggests that larger datasets or advanced models, such as GPT-4, could improve text analysis. SHAP results in Figure 5 emphasize the significance of infection and fever, indicating a need for refined feature extraction. Fairness metrics in Figure 4 reveal disparities that necessitate broader demographic analysis, aligning with Yadgarov et al.'s advocacy for standardized methods [5]. Future efforts will employ SMOTE to address data imbalances and focus on real-time ICU monitoring, drawing inspiration from Hu et al.'s COMPOSER model, to facilitate swifter interventions [6].

References

[1] M. Singer, C. S. Deutschman, and C. W. Seymour, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[2] C. J. Paoli, M. A. Reynolds, and M. Sinha, "Epidemiology and costs of sepsis in the united states," *Critical Care Medicine*, vol. 46, no. 12, pp. 1889–1897, 2018.

[3] A. Kumar, D. Roberts, and K. E. Wood, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[4] S. Zhou, Y. Wang, and X. Zhang, "Interpretable machine learning for predicting sepsis risk in emergency triage patients," *Scientific Reports*, 2025.

[5] M. Y. Yadgarov, J. Smith, and K. Brown, "Early detection of sepsis using machine learning algorithms: A systematic review and network meta-analysis," *Frontiers in Medicine*, 2024.

[6] C. Hu, Z. Liu, and Y. Jiang, "Impact of a deep learning sepsis prediction model on quality of care and survival," *npj Digital Medicine*, 2024.