Decoding Physiological Waveforms for Early Prediction of Sepsis

Nikhil Choudhary[1], Liora Dsilva[1], Yifan Ding[1], Naisargik Lenka[1], Eileen Su[1], Preethan Bachina[1], Carl Harris[1], Casey Overby-Taylor[1], Joseph L. Greenstein[1], Robert D. Stevens[1,2,3,4]

Departments of [1]Biomedical Engineering, [2]Anesthesiology and Critical Care Medicine, [3]Neurology, [4]Neurosurgery at Johns Hopkins University
EN.580.680: Precision Care Medicine

Introduction

- **Sepsis,** defined as a dysregulated host response to infection, is a leading cause of death globally with a mortality rate of **20-30%**.
- Early recognition and treatment of sepsis are vital; delays in antibiotic initiation in septic shock increase mortality by **7% per hour**.
- Existing sepsis prediction models—which rely on clinical variables, laboratory results, and non-waveform physiological data—have **limited accuracy** for sepsis-related outcomes of interest.
- **Physiological waveforms** (i.e., ECG, PPG, ABP) contain features which lend themselves to machine learning analysis and predictive modeling.
- Foundation models provide a powerful framework for waveform analysis, capturing complex temporal dependencies and physiological patterns with high fidelity.

Objectives

- The principal aim is to build a model for **sepsis onset and outcome prediction** that leverages high-frequency physiological waveform data.
- The core hypothesis is that physiological waveforms contain latent features that are markers of the early preclinical stages of sepsis, i.e. are predictive of sepsis.

Approach

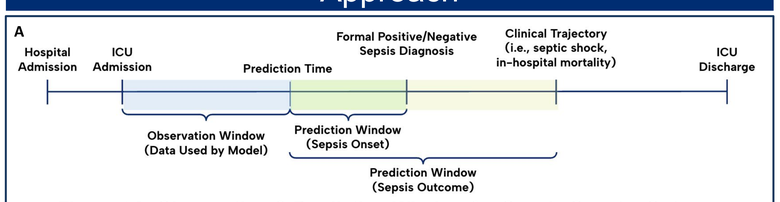


Figure 1A. Observation & Prediction Window for Sepsis Onset + Outcome

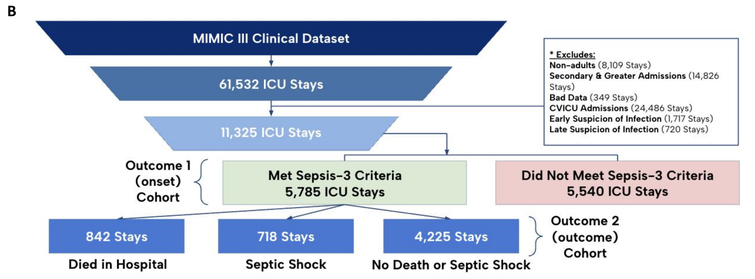


Figure 1B. Cohort Selection Criteria

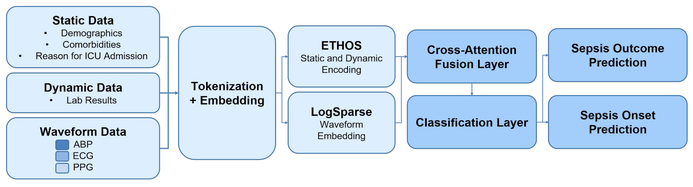


Figure 2. Proposed Model Pipeline

- Static and dynamic tokens are processed by **ETHOS**[1]—a transformer architecture with proven clinical utility in ICU-related tasks.
- Waveform tokens are processed in parallel using a modified version of ETHOS with **LogSparse self-attention** instead of full self-attention.
- A cross-attention fusion layer integrates embeddings from both streams.

Tokenization

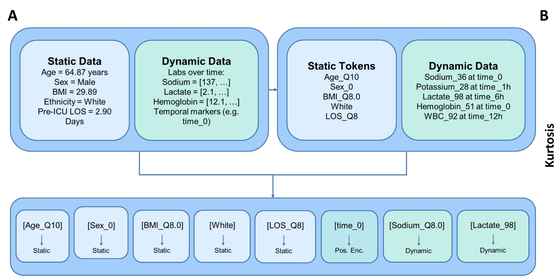


Figure 3A. Transformation of Raw EHR Data into Structured Token Sequences. Static tokens are prepended to each patient sequence, followed by dynamic tokens separated by time tokens to encode temporal gaps between measurements.

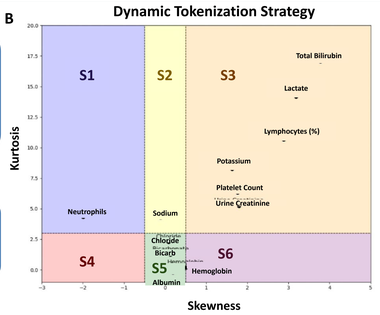


Figure 3B. Tokenization Strategy for Lab Features. Each lab is discretized into 8 tokens using one of six strategies (S1-S6), determined by the underlying distribution of values in patient cohorts.

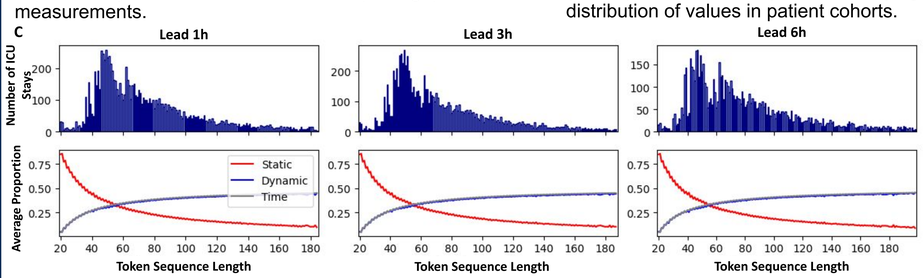


Figure 3C. Distribution (Top) and Composition (Bottom) of Token Sequences by Lead Time. Longer lead times (e.g., 6h) result in shorter token sequences (on average) due to limited data availability. As sequence length increases, the proportion of dynamic and time tokens dominates, while the proportion of static tokens as a function of total tokens decreases.

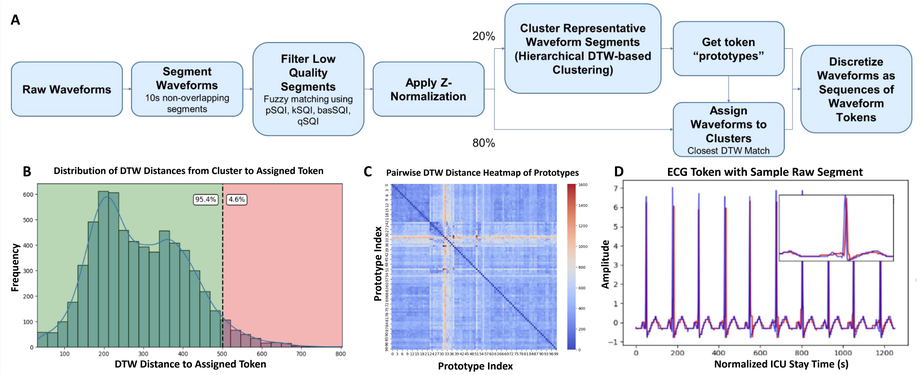


Figure 4. Waveform Tokenization. (A) Overview of pipeline for discretizing raw ECG waveforms into representative tokens. **(B–D)** Results from applying this method to 10,000 ten-second ECG segments clustered into 100 representative tokens. **(B)** Over 95% of segments matched well (DTW distance < 500) to a token prototype. **(C)** A majority of clusters were well-separated indicating distinct token prototypes. **(D)** Sample token prototype (red) overlaid with sample raw waveform segment (blue) assigned to that cluster.

Results

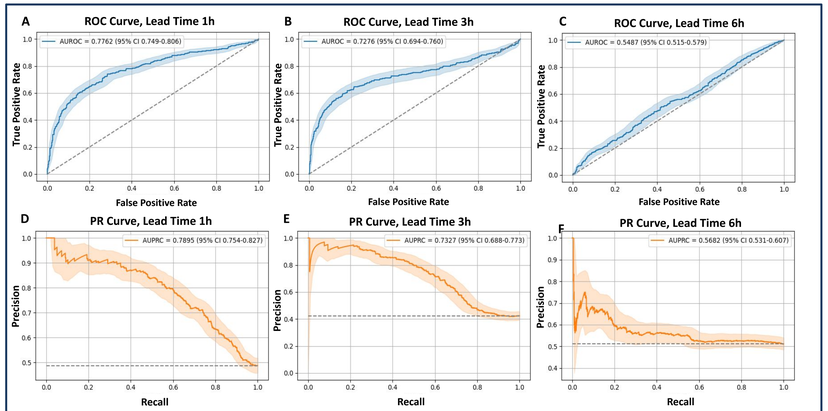
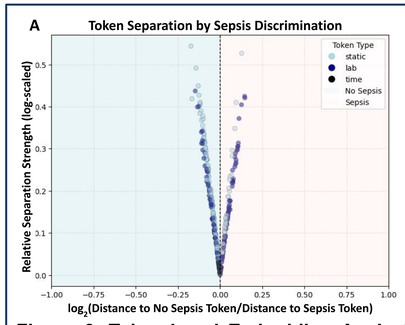


Figure 5. ROC and PR Curves for Sepsis Onset Prediction at 1h, 3h and 6h Lead Times (Static and Dynamic Tokens Only). Model performance is relatively strong at shorter lead times, with higher AUROC and AUPRC values observed at 1h and 3h. Performance deteriorates substantially at a 6h lead time, with both metrics approaching levels expected from a no-skill classifier.



B

Top 10 Tokens for Predicting Sepsis Outcome	
No Sepsis	**Sepsis**
• Chloride Q8 (max)	• Age (25-30)
• Magnesium Q1 (min)	• Age (16-20)
• Age (> 90) (max)	• Lactate Q2
• Neutrophils Q2	• Total Bilirubin Q6
• Neutrophils Q1 (min)	• RoA_ENT Q2

Figure 6. Token-Level Embedding Analysis of Sepsis Outcome Predictions. (A) The x-axis is the log-ratio of cosine similarity distances to the sepsis_yes and sepsis_no token embeddings (i.e., the classification tokens), while the y-axis indicates relative separation strength. Time tokens show minimal separation, while both static and lab-derived dynamic tokens exhibit predictive signal. **(B)** Top 10 tokens most proximal (by cosine similarity) to the sepsis and no-sepsis token embeddings. Extreme lab values and age groups seem more significant for outcome discrimination.

Conclusion

- Using only **static and dynamic clinical features**, we were able to predict sepsis onset with **moderate confidence**. However, these feature sets may fail to capture the **rich temporal and physiological complexity** underlying early sepsis.
- We hypothesize that incorporating **high-frequency physiological waveform data** will significantly improve model performance, especially at **longer lead times**, by enabling more precise detection of **subtle, preclinical changes** indicative of sepsis.
- As a next step, we will focus on **integrating waveform tokens** using the approach in **Fig. 2** and **Fig. 4A** to fully leverage the potential of the **transformer-based foundation model** to extract latent features and improve **early sepsis onset and outcome prediction**.

[1]Renc P, Jia Y, Samir AE, Was J, Li Q, Bates DW, Sitek A. Zero shot health trajectory prediction using transformer. NPJ Digit Med. 2024 Sep 19;7(1):256. doi: 10.1038/s41746-024-01235-0. PMID: 39300208; PMCID: PMC11412988.