# RARE EVENT EARLY DETECTION: A DATASET OF SEPSIS ONSET FOR CRITICALLY ILL TRAUMA PATIENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Sepsis is a major public health concern due to its high morbidity, mortality, and cost. The clinical outcome can be substantially improved through early detection and timely intervention. During the past decade, by leveraging publicly available datasets, machine learning (ML) has driven advances in both research and clinical practice. However, existing public datasets consider mainly the general ICU (Intensive Care Unit) population and neglect the difference that trauma patients are having. In critically ill trauma patients, injury-related inflammation and organ dysfunction can increase the risk of sepsis, while also masking the clinical signs of infection. Therefore, a targeted identification of post-traumatic sepsis is critical but challenging, which was rarely studied before. To address this gap, we introduce a publicly available standardized post-trauma sepsis onset dataset extracted, relabeled using standardized post-trauma clinical facts, and validated from MIMIC-III (a large database with more than 40,000 patients who stayed in critical care units). Furthermore, we frame our early detection problem of post-trauma sepsis onset according to the ICU's real running routine that was ignored before, which results in a daily sepsis onset early detection problem for each patient, and sepsis onset becomes a rare event. In this work, we also establish a general benchmark to address this rare event challenge through comprehensive experiments, which shows the necessity of further advancements using this new dataset for early detection of rare events. The code for dataset extraction and usage is available at `https://github.com/ML4Health-AnonResearch/PostTraumaticSepsis.git`.

## 1 INTRODUCTION

Sepsis is a life-threatening disease associated with an altered immune response to infection and remains a major public health concern with high morbidity, mortality, and economic burden (Singer et al., 2016; Fleischmann et al., 2016; Rudd et al., 2020). To leverage rich electronic health records (EHRs) and advanced ML techniques for sepsis management, the AI Clinician (Komorowski et al., 2018) pioneered ML-driven optimization for sepsis treatment by introducing a structured and reproducible approach to generate sepsis cohorts from public ICU datasets. This work not only fostered collaborations between medical and ML researchers, but also laid the foundation for subsequent large-scale sepsis studies. Later, recognizing the importance of early intervention for sepsis, PhysioNet launched a challenge focused on early detection of sepsis, attracting widespread engagement from academia and industry (Reyna et al., 2020). The dataset from this challenge continues to influence ongoing research by providing a common benchmark for fair comparison across methods. These efforts illustrate how public datasets have not only facilitated ML research on sepsis but also demonstrated impact in real-world clinical care. Notably, the Targeted Real-Time Early Warning System (TREWS), an ML-based alert system implemented across multiple hospitals in the United States, demonstrated how ML can improve sepsis management in clinical settings (Adams et al., 2022).

Despite the overall improvement in general outcomes of sepsis in the ICU, hospital-acquired sepsis remains a prevalent complication and a significant contributor to morbidity and mortality after severe traumatic injury (Stern et al., 2023; Guirgis et al., 2016). This could be caused by the delay in the detection of post-traumatic sepsis. This delay is mainly due to the fact that critically ill trauma patients often exhibit physiological responses and alterations in organ function (e.g., acute lung injury and acute kidney injury) that are due to the initial trauma and exist before the development of infection (Eguia et al., 2020; Eriksson et al., 2019; Stern et al., 2023). This unique pathophysiological overlap makes the trauma cohort fundamentally different from the general ICU population, requiring tailored definition criteria rather than directly applying the definition of existing ICU-based standards.

Despite the fact that targeted identification of post-traumatic sepsis is clinically necessary, this has rarely been studied before, especially in the ML community. Public datasets underlying the AI Clinician (Komorowski et al., 2018) and the PhysioNet Challenge (Reyna et al., 2020) both focus on the general ICU population, and therefore only provide

sepsis labels based on the standardized Sepsis-3 definition (Singer et al., 2016). For convenience, research groups studying trauma cohorts (Guo et al., 2024; Fu et al., 2019) directly use readily available sepsis-3 labels, without incorporating trauma-specific clinical guidance. As a result, current ML researchers often rely on datasets that do not accurately reflect the realities of trauma care. To address this gap, we introduce a publicly available standardized post-trauma sepsis onset dataset that is extracted, relabeled, and validated from MIMIC-III, a large database with more than 40,000 patients who stayed in critical care units. Specifically, our dataset is the first to provide a well-defined standardized trauma cohort with reliable post-traumatic sepsis onset labels. The dataset is supported by a well-documented reproducible code base for future research and clinical applications.

Furthermore, guided by real clinical practice in the ICU for trauma patients, we frame the early detection problem of sepsis onset as a daily detection problem for each patient, since each patient is comprehensively discussed only once in the early morning for daily treatment. This makes the onset of sepsis a rare event. Early detection of rare events is a fundamental challenge in ML, not limited to sepsis, making our dataset a valuable benchmark for developing and evaluating such methods. To support this research, we conducted a comprehensive empirical study and established a general benchmark on this rare event early detection problem, highlighting future opportunities for improvement. Our contributions can be summarized as follows.

- We are the first to emphasize the difference between general ICU sepsis and post-traumatic sepsis, and to build a standardized trauma cohort with targeted post-traumatic sepsis onset labels to facilitate early detection of post-traumatic sepsis onset.
- We frame early detection of post-traumatic sepsis onset according to real clinical practice in the ICU, which has been ignored before and results in a rare event early detection problem.
- We conducted comprehensive experiments and established a general benchmark for early detection of post-traumatic sepsis onset, showing future opportunities using this new dataset.

## 2 Related Work

This section reviews related work from two perspectives. We first examine sepsis-related datasets, with emphasis on how sepsis is defined, how detection tasks are framed, and how study cohorts are extracted, discussing their reliability, reproducibility, and suitability for the ICU trauma population. We then discuss machine learning techniques for early sepsis detection and broader approaches for handling rare event detection.

### 2.1 Publicly Available Sepsis Definition

For a sepsis dataset, it is critical to locate sepsis patients and the timestamp of sepsis onset. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) (Singer et al., 2016) is widely used and is the current consensus definition of sepsis in the ML community. AI Clinician (Komorowski et al., 2018) pioneered the use of machine learning for sepsis-relevant tasks. By leveraging public datasets such as MIMIC-III (Johnson et al., 2016) and eICU (Pollard et al., 2018), AI Clinician provides a structured framework for sepsis definitions based on the Sepsis-3 criteria. Publicly available datasets and their corresponding open-source code bases ensure the reproducibility of Sepsis-3 and facilitate future sepsis-related research. Later on, the PhysioNet Challenge 2019 (Reyna et al., 2020) introduced a standardized benchmark for early sepsis detection. Although it provides sepsis labels for each patient along with documentation on how sepsis was defined, the reproducibility of sepsis labels for future remains limited. In addition, the official MIMIC code repository has recently released SQL implementations of Sepsis-3 (for Computational Physiology, 2025), further supporting reproducibility and consistency in sepsis-related studies.

Although all above public datasets follow the general Sepsis-3 guideline, describing sepsis as life-threatening organ dysfunction resulting from a dysregulated host response to infection, their detailed criteria are not identical. For example, AI Clinician identifies organ dysfunction when the SOFA score is greater than 2, while PhysioNet defines it based on a SOFA score increase of 2 (Johnson et al., 2018). Moreover, PhysioNet enforced stricter feature extraction by emphasizing blood cultures and intravenous antibiotic routes. As a result, AI Clinician tends to identify a larger, less critically ill cohort, resulting in higher false positive rates. While PhysioNet improves, its reproducibility and clarity of the sepsis-3 label is limited without open source code support.

Beyond the variations in how Sepsis-3 criteria are operated across studies, another limitation is that the Sepsis-3 definition itself does not fully address the unique characteristics of trauma patients. Despite the importance of task-specific post-trauma sepsis definitions, currently there are no publicly available datasets addressing this. (Stern et al., 2023) provides a medical post-trauma sepsis guideline, yet no publicly released dataset has incorporated this more accurate definition. Given the high morbidity and mortality of hospital-acquired sepsis among trauma patients, recent

studies have explored sepsis detection in this subgroup of patients in the ICU. However, existing trauma cohort studies either rely on private datasets (Stewart et al., 2023; Ewig et al., 2023a; Guo et al., 2024) or use public datasets but with diverse cohort extraction criteria, and lack open source code support (Li et al., 2023; Fu et al., 2019). This severely constrains the reproducibility of trauma-specific studies. In this work, we adopt a refined trauma-specific sepsis definition on a public dataset with a reproducible code base, ensuring a standardized foundation for future research of post-traumatic sepsis.

### 2.2 Detection Task Framing

As discussed in (Stewart et al., 2023), most previous studies used a fixed time to onset framing, limited to where the time of onset is known a priori. The other portion of previous studies generates detection results at regular intervals using all data available from admission, where the variable observation window is a major challenge. More importantly, they did not consider the real clinical practice needs. In this work, we focus on whether sepsis onset will occur on a daily basis. We time delivery of daily detection for the early morning, right before the routine morning rounds, so that medical staff can apply this information to patient care decisions throughout the day.

### 2.3 Early Sepsis Detection

Early sepsis detection involves addressing two main challenges. The first is how to derive meaningful physiological representations from raw, noisy time-series data within an observation window. The second is how to distinguish early sign of sepsis cases from all cases based on the given features.

For the challenge of feature representation, a straightforward solution is feature engineering, where handcrafted statistical descriptors such as mean, maximum, or minimum are computed over each vital sign within the observation window (Ewig et al., 2023a; Li et al., 2019). Such features provide a coarse summary of the patient's physiological state but fail to capture temporal dynamics. To address this limitation, (Morrill et al., 2019) introduces mathematical tools such as the signature transformation (also known as path signature), which converts multivariate time series into fixed-length feature vectors that preserve temporal information. More recently, deep learning models have been widely adopted to automatically extract hidden temporal features, with commonly used architectures including recurrent networks such as gated recurrent units(GRUs) (Li et al., 2019; Nonaka & Seita, 2019) and Long-Short-Term Memory(LSTMs) (Liu et al., 2019; Stewart et al., 2023; Elmerahi et al., 2024; Ramos et al., 2021), as well as temporal convolutional neural networks (TCNNs) (Lee et al., 2024; Lauritsen et al., 2020b). It should be noted that although deep learning models have greater potential to learn complex features, they have a high demand for the dataset size.

Given the feature representations, the classification can be broadly grouped into two categories: tree-based models and fully connected layer (FCL) classification heads. Tree-based models, such as XGBoost and LightGBM, have been widely applied in early sepsis detection (Morrill et al., 2019; Ewig et al., 2023a). They are friendly to small datasets and offer fast training, but their performance heavily depends on the quality of the input features, and they are limited in capturing complex, high-dimensional relationships. In contrast, FCL classification heads are typically used on top of deep representation models (Stewart et al., 2023; Tran et al., 2019). They possess a stronger expressive power, allowing them to capture non-linear interactions among features. However, they require larger training datasets with high-quality labels and are prone to overfitting.

### 2.4 Imbalance and Rare Event Challenge

Like most disease detection tasks, early sepsis detection also faces an imbalance challenge or even a rare event challenge. To mitigate the imbalance, most of the work applies data resampling techniques, such as oversampling and undersampling (Lauritsen et al., 2020a; Teredesai et al., 2022; Ewig et al., 2023b). However, these methods often lead to overfitting, in the case of oversampling, or loss of critical information, in the case of undersampling. In contrast to this, reweighting-based strategies have also been employed to alleviate the imbalance by assigning different loss weights to positive and negative samples (e.g., inversely proportional to class frequency) (Elmerahi et al., 2024; Zhang et al., 2024). However, reweighting struggles with highly imbalanced datasets as extreme weight differences can lead to unstable training and overfitting to the minority class.

For these reasons, conventional imbalance handling methods remain insufficient, especially in rare event detection tasks like ours, where the minority class not only constitutes a small fraction but also has a limited number of samples. As a result, some have proposed to use representation learning to help address the challenges posed by class imbalance. Ramos et al. (2021) proposed a novel approach using unsupervised anomaly detection. They first apply an autoencoder to learn feature representations, and then use a clustering algorithm to classify samples into normal and anomalous groups. Although such methods can help identify rare events, they rely on the assumption that anomaly

samples have a distinct distribution, which does not always hold true. In our case, trauma patients with sepsis often exhibit physiological patterns similar to general trauma patients, making simple anomaly detection less effective. Alternatively, (Stewart et al., 2023) utilizes self-supervised pre-training to help alleviate class imbalance.

There are other methods to address class imbalance or rare events in other domains. Some methods conduct minor class data oversampling via synthetic data generation, the most notable of which is the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which has been widely adopted, though not specifically for sepsis. In this work, we will conduct a comprehensive empirical study to explore their potential in handling the rare event early detection for post-traumatic sepsis onset.

## 3 Dataset Construction

Using the MIMIC-III dataset (Johnson et al., 2016), an open-access anonymized database of 61,532 admissions from 2001 to 2012 across six ICUs, our work aims to construct a public dataset to facilitate early post-traumatic sepsis onset detection. We are the first to extract and publish a dataset that incorporates updated clinical insights on sepsis targeting trauma cohorts. It consists of three modules: standardizing a trauma-focused cohort, adopting carefully designed sepsis criteria targeting trauma patients based on the MIMIC-III dataset, and introducing a clinically aligned detection frame. The selection of critical criteria will be defined, explained, and referenced in this work to ensure the quality of the dataset. Additional details are provided in the Appendix and our GitHub repository to facilitate dataset reconstruction and support future research on this dataset.

### 3.1 Cohort Extraction: Critically Ill Trauma Patients

The critically ill trauma cohort is defined using standard trauma cohort selection criteria, similar to those employed in previous studies (Li et al., 2023; Fu et al., 2019), with extraction criteria targeting critically ill patients following the guidance of (Stern et al., 2023). The extraction process starts with the identification of patients with valid data. Specifically, valid patients are defined as those hospital admissions (HADM_ID) that include at least one ICU stay and corresponding data in CHARTEVENTS, which serves as the primary repository for a patient's information during their hospital stay, encompassing vital signs, laboratory values, and ventilator settings.

We then selected admissions associated with traumatic injuries based on a carefully curated list of ICD-9 E-codes that broadly represent trauma-related events. This list excludes categories such as poisoning to ensure relevance and consistency with trauma-specific cohorts. The full list of included codes is provided in Appendix B.1.

We then refined our study cohort to include only adult patients aged between 18 and 89 years, with an admission duration of 48 hours or more. This age range is selected because the lower limit of 18 represents adult patients, consistent with standard definitions in medical research (Komorowski et al., 2018; Li et al., 2023). The upper limit of 89 is established because, in the MIMIC-III dataset, patients older than 89 are recorded as being 300 years old to protect their privacy (Johnson et al., 2016). The admission duration criterion is designed to exclude patients who are at low risk for the development of sepsis, either because they died or recovered quickly.

Finally, we included only patients with three or more days of mechanical ventilation to identify critically ill patients at higher risk of developing sepsis. Ventilation days are defined as the total number of days that each patient received mechanical ventilation during a single admission, regardless of the number of hours on each day. Although this criterion is not typically included in trauma cohort extractions, it is crucial for identifying patients with a higher likelihood of sepsis, as noted by (Stern et al., 2023; Horn et al., 2022). This trend is evident in the MIMIC-III dataset, where the sepsis ratio shows a rapid increase starting from patients with three days of ventilation, as illustrated in Fig. 2 in the Appendix.

In the MIMIC dataset, the care unit defines the type of ICU, such as the Coronary Care Unit (CCU), Trauma/Surgical ICU (TSICU), or Neonatal ICU (NICU). While many hospital admissions involve only one ICU stay, patients may transfer between these specialized ICUs based on their medical needs during a single hospital stay. Additionally, a patient can be admitted to the hospital for an extended period (e.g., a month) before being transferred to an ICU. Unlike (Komorowski et al., 2018), which used ICUSTAY_ID, we selected HADM_ID (hospital admission ID) as the instance ID for our cohort. HADM_ID encapsulates all medical interventions and observations during a specific hospital admission, aligning with the concept of a "patient" in the clinical research of sepsis studies. This also aligns with how ICD-9 diagnosis codes and billing are assigned based on the entire hospital stay rather than individual ICU stays. This approach allows for comprehensive tracking of patient outcomes, providing a more complete view of the entire hospital stay. Subsequently, we gathered a refined cohort of critically ill trauma patients suitable for our study

on early sepsis onset detection, comprising a total of 1,570 admissions, as summarized in Fig. 3 in the Appendix, where the detailed patient characteristics are shown in Table 4.

### 3.2 Post-trauma Sepsis Definition

Our work here relies on a clinical post-trauma sepsis identification defined by (Stern et al., 2023), adapted from the Centers for Disease Control and Prevention's adult sepsis surveillance criteria and the original Sepsis-3 guidelines. We define hospital-acquired post-traumatic sepsis within the MIMIC-III dataset as a clinically suspected infection accompanied by acute organ dysfunction, with modifications tailored to the unique challenges of classifying sepsis in our trauma cohort. This methodology involves two key steps: **Pre-processing feature tables**, focusing on extracting and preprocessing relevant data from MIMIC-III to identify post-trauma sepsis cases through careful feature selection and qualified record processing based on cross-referenced multi-source data; and **Post-trauma Sepsis onset criteria**, where we establish specific criteria for sepsis onset targeted at trauma patients. In the initial phase, we preprocessed pertinent data from three primary tables: blood culture, antibiotics, and a modified version of the Sequential Organ Failure Assessment (SOFA) score. The blood culture and antibiotic data are jointly analyzed to identify suspected infections, while the SOFA score is used to define the onset and quantify the severity of organ dysfunction.

#### 3.2.1 Pre-processing: blood culture

For blood cultures, we apply a filter to capture relevant entries occurring at or after 72 hospital hours. We focus exclusively on blood culture criteria, in contrast to the approaches taken by AI Clinician works (Komorowski et al., 2018) and (Stern et al., 2023), which utilize all body tissue cultures (e.g., blood, urine, sputum). Following the methodology of (Rhee et al., 2019; Reyna et al., 2020), we emphasize blood cultures because they are typically part of a panel of samples collected when sepsis is suspected. When clinicians suspect an infection and potential sepsis but are uncertain about the source, blood cultures are obtained alongside other body fluid samples. Unlike other cultures (such as tracheal or urine cultures), which may be collected for surveillance of antibiotic-resistant organisms and might not indicate a suspected infection, blood cultures specifically aim to identify systemic infections. While this approach may result in a lower number of identified infections and sepsis cases compared to using all body fluid cultures, it is likely more specific for diagnosing systemic infections critical for sepsis identification. The filter for blood cultures at or after 72 hospital hours excludes cases of sepsis acquired before hospital admission (Stern et al., 2023; Rhee & Klompas, 2020).

#### 3.2.2 Pre-processing: antibiotic

Preprocessing antibiotic data from the Prescriptions table in the MIMIC dataset is complicated by the lack of standardized antibiotic labels and the disordered nature of the data, which often includes overlapping or fragmented entries and inconsistent drug name formats. Most previous work has overlooked these significant challenges, making it difficult to study and reproduce sepsis-related datasets using MIMIC. This lack of transparency leads to two major drawbacks: first, different processing methods for antibiotics can result in inaccurate sepsis identification, such as including antibiotics administered for prophylactic purposes, which may increase false positive cases; second, from a machine learning perspective, this inconsistency hampers the reproducibility of the dataset, reducing comparability when training methods are applied across different studies.

To address this gap, we propose a detailed, post-trauma sepsis-specific set of criteria for extracting antibiotic records, following the guidance of (Stern et al., 2023), to ensure the accuracy and relevance of the data for defining sepsis onset in trauma patients. The criteria begin with identifying qualifying antibiotic drug names, selected based on cross-referencing multiple sources (Stern et al., 2023; Johnson et al., 2016; Komorowski et al., 2018) tailored specifically for the treatment of trauma-related sepsis. Further details are provided in Appendix C.1. We restrict the criteria to all intravenous (IV) antibiotics and two designated oral antibiotics—vancomycin and linezolid—while excluding prophylactic antibiotics and those administered on the first day. We also ensure that the same qualifying antibiotic is not administered within the previous day to identify new antibiotic orders accurately. We also require that a qualifying antibiotic being administered for a minimum of four consecutive days or until the patient's death or discharge, without necessitating the same antibiotic throughout this period (Stern et al., 2023; Rhee et al., 2019). To ensure the accuracy and relevance of the data for defining hospital-acquired sepsis, the final qualified antibiotic events must meet all of the above criteria and only the starting time of the coherent antibiotic events will be used to identify the suspected infections. Less-critical but necessary implementation details can be found in our GitHub project, which aims to streamline the extraction of antibiotic orders and transform them into coherent events, thereby guaranteeing the reproducibility of the dataset.

#### 3.2.3 PRE-PROCESSING: MODIFIED SOFA SCORE

Concurrently, we used a modified SOFA score targeted for trauma cohort (Stern et al., 2023; Rhee et al., 2019; Bosch et al., 2022), which excludes the urine output (uo) and Glasgow Coma Scale (GCS) variables utilized in the traditional SOFA score calculation. Excluding the Glasgow Coma Scale (GCS) from the SOFA score calculation for trauma patients is due to the confounding influence of traumatic brain injury, which can significantly alter GCS values irrespective of organ failure. This approach aligns with historical practices (Minei et al., 2012; Horn et al., 2022) in trauma-specific organ failure scores and helps ensure that neurological impairments caused by trauma do not skew the assessment of other organ systems. Additionally, we omitted urine output in the renal component as per a validated modification targeted on critically ill patients.

#### 3.2.4 POST-TRAUMA SEPSIS-3 ONSET CRITERIA

In the subsequent phase, we define sepsis patients based on the pre-processed feature tables. For each patient, we specify the following three time points to determine the onset time of sepsis $t_{sepsis}$ as illustrated in Fig. 5 of the Appendix.

- $t_{infection}$: Clinical suspicion of infection, identified as the timestamp when a culture is ordered within a 5-day window of antibiotic initiation for qualified processed antibiotic events (Stern et al., 2023; Rhee et al., 2019).
- $t_{SOFA}$: Occurrence of organ failure, as identified by at least a 2-point increase in the modified SOFA score within a 7-day window, which includes 3 days before, the day of, and 3 days after the qualifying culture, following the guidance of (Stern et al., 2023; Reyna et al., 2020).
- $t_{sepsis}$: Onset of sepsis, identified as the earliest culture timestamp that meets the criteria for both suspicion of infection and organ failure. Since early sepsis detection focuses solely on the first onset event, this timestamp is crucial for timely intervention.

Among the 1,570 trauma admissions, we identified 729 admissions with potential infections and 535 patients with sepsis. The distribution of onset days for sepsis is depicted in Fig. 4 in the Appendix. Notably, the peak for culture orders, indicating the onset of sepsis, occurs on day 5, which aligns with clinical experience (Horn et al., 2022).

### 3.3 EARLY SEPSIS DETECTION SETUP

In our dataset, we adopt a deployable daily detection setup that is closely aligned with ICU workflows. Predictions are generated each morning, just before routine rounds, so that medical staff can incorporate the results into patient care decisions for the upcoming day. Specifically, data collected during the previous night are used to estimate the likelihood of sepsis onset within the next 24 hours. Beyond its clinical alignment, this setup also improves data quality, as nighttime records are generally less affected by external interventions (e.g., surgeries and diagnostic procedures).

#### 3.3.1 FEATURE EXTRACTION

The features collected during nighttime hours (from 6:00 p.m. to 6:59 a.m. the following day) are used for early sepsis detection within the next 24 hours. We extracted seven key vital sign features: heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, temperature, and SpO2. These features are critical indicators of physiological status, which can be helpful in the early detection of sepsis. While the current feature set is sufficient for early sepsis onset detection, additional features can be extracted. Our repository will be updated and maintained to include more features, such as cumulative exposures and lab results, in the future.

#### 3.3.2 INSTANCE CONSTRUCTION

The raw data is first aggregated into average hourly records, converting it into a 2D time-series format with a shape of $(T, F)$, where $T = 13$ represents the 13 hourly timestamps over a night (6 p.m., 7 p.m., ..., 6 a.m.), and $F = 7$ represents the seven key vital sign features. We then filter the nights to include only those from day 2 to day 14 since the patient's admission, focusing on the critical period for early sepsis detection. Infections occurring before day 2 are considered to have been acquired before hospital admission and are therefore excluded from our study. Additionally, records beyond day 14 are excluded, as patients are classified as having chronic critical illness after this period, which presents a different phenotype than the post-trauma sepsis we are targeting for detection.

To ensure compatibility with most machine learning approaches, we provide a standard dataset without missing values, referred to as the "S dataset". Missing values were imputed using a forward and then backward filling method within

a window from 7:00 a.m. (right after the end of the previous night) to 6:59 a.m. (before the end of the current night). Although backward filling is applied, this approach remains deployable as it does not rely on timestamps beyond the nighttime period. Instances that still contain missing values after this process are removed. We also include the "N dataset", which retains NaN values, to preserve raw data without introducing imputation bias, as detailed in Appendix D.

These nighttime records were then assigned 0/1 labels based on the patient's sepsis onset time. If sepsis onset occurs within 24 hours after 6:59 a.m. of the corresponding record, it is labeled as positive; otherwise, it is labeled as negative. This means all nighttime instances of non-sepsis patients will be labeled as negative. For sepsis patients, only the night immediately preceding the sepsis onset will be labeled as positive. For example, if a nighttime instance starts at 10:00 p.m. on day $i$ and ends at 6:59 a.m. on day $(i+1)$, this instance will be labeled as positive only if the sepsis onset occurs within the window from 7:00 a.m. on day $(i+1)$ to 6:59 a.m. on day $(i+2)$. All nighttime instances of the sepsis patient before the positive instance will be labeled negative, and all nighttime instances following the positive instance will be discarded, since they are not relevant to early detection.

To demonstrate that these vital sign features can show early signs of sepsis, we visualize the temporal trends of key physiological features (e.g., heart rate, temperature) in the four days leading up to sepsis onset, using Accumulated Days Before Sepsis Onset (DBSO) on the x-axis. Fig. 1 highlights noticeable changes, particularly in temperature and heart rate at x = -1, suggesting that the delta values between consecutive nights may enhance early sepsis detection. A full visualization of all physiological features is provided in Appendix C.4.

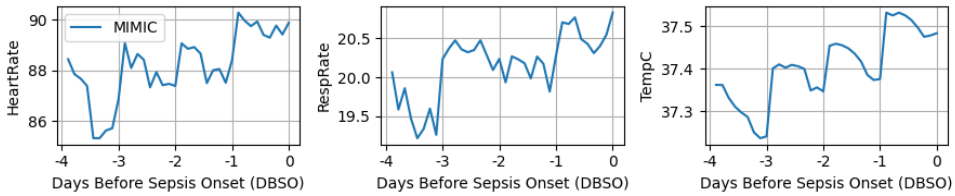


Figure 1: Subset of physiological trends in the four days before sepsis onset (x = 0 marking sepsis onset).

Finally, our post-trauma early sepsis detection data includes 440 positive and 8,319 negative instances in the S dataset (8,759 cases across 1,522 unique patients), resulting in an imbalance ratio of 0.05 (positive/all instances). The N dataset contains 455 positive and 8,522 negative instances (8,977 cases across 1,535 unique patients), with an imbalance ratio of 0.051. The study population is smaller than the original trauma cohort (1,570 patients) due to missing data in MIMIC, and drop details are explained in Appendix D. These missing EHR data may stem from sensor or equipment malfunctions, irregular documentation practices, or incomplete data storage. In the experiments, we will demonstrate the importance of these reliable labels on a relatively small subset of qualified data.

## 4 Benchmark Task & Baselines

Building on the dataset described in the previous section, we now turn to the benchmark task and baseline methods, with a focus on the rare event detection challenge introduced by the night-time framing.

### 4.1 Task Definition

The rare event early detection task supported by this dataset is the early detection of sepsis onset in trauma patients using nighttime vital signs. Specifically, we formulate this problem as a binary classification problem. Let $\{(x_i, x_i^{\Delta}), y_i\}_{i=1}^{n}$ denote the dataset, where $x_i$ represents the nighttime vital signs, $x_i^{\Delta}$ represents the change (delta) between the current night's data $x_i$ and its previous night's data as inspired by the observations in Fig. 1, and $y_i$ is the sepsis label. To address the severe class imbalance problem, we propose to do reconstruction-based representation learning. Thereafter, to capture more local variance for sepsis instances, we can apply random masks and generate synthetic data using the pre-trained reconstruct model. In the following, we describe the model architecture followed by the detailed training procedures.

### 4.2 Model Architecture

**Encoder** The encoder processes time-series data using a combination of GRU and convolutional layers to capture both temporal patterns and feature correlations. It begins with a Bidirectional GRU layer that learns temporal de-

pendencies in both directions, followed by another GRU layer to refine the feature representation. Next, the model applies two 1D Convolutional layers, which further extract local feature correlations. To improve stability and prevent overfitting, kernel regularization, Batch Normalization, and Dropout layers are applied after both the recurrent and convolutional layers. Finally, a Global Average Pooling layer compresses the extracted features into a compact vector while preserving the overall feature distribution.

**Autoencoder** To learn the distinct distributions from two input sources—current night data and delta data—we adopt a multi-modal encoder architecture. This consists of two separate encoders, one for each input, ensuring that each modality is processed independently before feature fusion. Each input is first passed through its respective encoder, where GRU and CNN layers extract meaningful temporal and local patterns. The encoded embeddings from both encoders are then concatenated to form a unified latent representation, which serves as input to the decoder. The decoder mirrors the encoder structure, following a symmetric design, where GRU units decrease in the encoder and increase in the decoder. This self-supervised reconstruction-based Autoencoder is designed to learn the input data distribution before fine-tuning, providing a strong initialization for downstream classification tasks.

**Classifier** The classifier takes the encoded feature representation from the pre-trained encoder and applies a fully connected classification head to predict the sepsis outcome. The input is first passed through the encoder, which extracts night-time feature representations. The encoded embedding is then processed by a Layer Normalization layer to stabilize training, followed by a Dropout layer to improve generalization and prevent overfitting. Finally, a Dense output layer with a softmax activation is used to generate a probability distribution over the two classes (non-sepsis and sepsis), making it suitable for binary classification.

### 4.3 Training Pipeline

**Stage 1: masked autoencoder (MAE) for feature learning** We adopt a Masked Autoencoder (MAE) to learn robust physiological representations of patients. We explore two pre-training settings: (1) on the trauma cohort itself, and (2) on a broader ICU population. Leveraging large-scale ICU data enables deep learning to learn stronger feature representations and alleviates the class imbalance challenge of rare sepsis events. When pre-training only on the trauma cohort, the masking strategy plays a dual role: enhancing feature learning and acting as data augmentation. After oversampling the minority class (sepsis), masking introduces local variability by reconstructing missing portions, encouraging the model to learn discriminative and resilient representations.

To further emphasize masked regions, we adopt a time-series targeted masking scheme of (Zerveas et al., 2021), which generates contiguous segments of masked values for each variable with lengths drawn from a geometric distribution, and apply a weighted Mean Absolute Error loss that prioritizes masked points while down-weighting unmasked ones. The weighting mechanism is defined as $mask\_weight = mask \times (1-\lambda) + (1-mask) \times \lambda$, where $\lambda$ is a hyperparameter controlling the weighting balance, set to 0.8 by default.

**Stage 2: classification with augmented representations** In Stage 2, classification is performed on an oversampled dataset, but unlike in Stage 1, where the masked inputs were directly used as augmentation, we now leverage the reconstructed outputs from the MAE model to generate augmented training samples. Since the masked-out regions are no longer directly supervised, this approach helps preserve local variance while generating more diverse representations of the minority class. The pre-trained MAE encoder is extracted and used as initialization for the classifier, allowing the model to leverage learned feature representations for better generalization. The classification model is then trained using binary cross-entropy loss on a balanced and augmented dataset, ensuring improved performance in detecting rare sepsis events.

### 4.4 Experiments

| | tn | fp | fn | tp | sensitivity | specificity | precision | f1_score | PR_auc | ROC_auc |
|---|---|---|---|---|---|---|---|---|---|---|
| Sepsis3 - NoPre | 0 | 1320 | 0 | 82 | 1.0 | 0.0 | 0.058488 | 0.110512 | 0.029244 | 0.499242 |
| Sepsis3 - GPre | 528 | 792 | 35 | 47 | 0.573171 | 0.4 | 0.056019 | 0.102063 | 0.063371 | 0.49757 |
| XGBoost | 1407.00 (36.71) | 256.80 (20.80) | 75.00 (5.34) | 13.00 (3.32) | 0.15 (0.04) | 0.85 (0.01) | 0.05 (0.01) | 0.07 (0.02) | 0.05 (0.01) | 0.53 (0.03) |
| LightGBM* | 1126.00 (120.84) | 537.80 (109.36) | 54.80 (9.55) | 33.20 (7.16) | 0.38 (0.09) | 0.68 (0.07) | 0.06 (0.00) | 0.10 (0.01) | 0.06 (0.00) | 0.55 (0.02) |
| GRU-TCNN - NoPre* | 1018.40 (112.54) | 327.80 (95.05) | 58.80 (8.17) | 23.20 (7.01) | 0.28 (0.09) | 0.76 (0.07) | 0.07 (0.02) | 0.11 (0.02) | 0.07 (0.01) | 0.55 (0.05) |
| GRU-TCNN - TPre* | 665.20 (66.98) | 681.00 (83.39) | 32.80 (11.30) | 49.20 (14.65) | 0.59 (0.16) | 0.49 (0.05) | 0.07 (0.01) | 0.12 (0.03) | 0.07 (0.01) | 0.58 (0.04) |
| GRU-TCNN - GPre* | 723.20 (62.81) | 623.00 (60.11) | 31.00 (3.00) | 51.00 (6.78) | 0.62 (0.05) | 0.54 (0.04) | 0.08 (0.01) | 0.14 (0.02) | 0.08 (0.01) | 0.60 (0.03) |

Table 1: Baseline Performance ( NoPre = No Pretraining; TPre = Pretrained on trauma cohort; GPre = Pretrained on general ICU cohort; * indicates training was conducted using Post-trauma Sepsis labels.)

In Table 1, the first two rows (Sepsis3-NoPre and Sepsis3-GPre) report baseline classification results obtained by training on the publicly available Sepsis-3 labels derived from the general ICU cohort and evaluating on our trauma

cohort with post-trauma sepsis labels. Because the general ICU dataset is sufficiently large, we did not apply 5-fold cross-validation for these two settings. For the rest of the experiments (XGBoost, LightGBM, GRU variants), models were trained/fine-tuned on the trauma-only dataset and evaluated using 5-fold cross-validation. To address the severe class imbalance, all experiments were trained on a balanced dataset obtained by randomly oversampling the minority (sepsis) class until it matched the size of the majority class.

First, we can observe that baselines without pre-training show very biased performance towards one of these two classes, indicating that simple oversampling cannot alleviate the severe class imbalance problem well. On the other hand, all baselines with pre-training provide better balanced results between the major and minor classes, which demonstrates the importance of pre-training for severe class imbalance. Then, we can observe that fine-tuning using way smaller amount of reliable post-traumatic sepsis labels compared to the large amount of general ICU sepsis labels, provides significantly better performance, which demonstrates the contribution of this new dataset. Finally, it is expected that pre-training using larger general ICU population can be more helpful.

| | pretrain | freeze_encoders | tn | fp | fn | tp | sensitivity | specificity | precision | f1_score | PR_auc | ROC_auc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SMOTE | NO | FALSE | 993.60 (110.65) | 352.60 (99.64) | 52.80 (8.07) | 29.20 (9.42) | 0.35 (0.10) | 0.74 (0.07) | 0.08 (0.01) | 0.12 (0.02) | 0.07 (0.01) | 0.57 (0.01) |
| Time Warp | NO | FALSE | 564.40 (303.10) | 781.80 (318.49) | 28.00 (21.52) | 54.00 (20.41) | 0.66 (0.26) | 0.42 (0.23) | 0.07 (0.01) | 0.12 (0.02) | 0.06 (0.01) | 0.55 (0.03) |
| Noise | NO | FALSE | 337.20 (177.11) | 1009.00 (176.41) | 14.80 (9.68) | 67.20 (5.93) | 0.82 (0.11) | 0.25 (0.13) | 0.06 (0.01) | 0.12 (0.01) | 0.07 (0.02) | 0.56 (0.05) |
| SMOTE | general icu | TRUE | 827.60 (56.73) | 518.60 (73.49) | 42.20 (5.26) | 39.80 (8.50) | 0.48 (0.09) | 0.62 (0.05) | 0.07 (0.01) | 0.12 (0.02) | 0.07 (0.01) | 0.57 (0.04) |
| Time Warp | general icu | TRUE | 735.20 (112.43) | 611.00 (136.21) | 35.20 (7.36) | 46.80 (7.05) | 0.57 (0.08) | 0.55 (0.09) | 0.07 (0.01) | 0.13 (0.02) | 0.08 (0.02) | 0.58 (0.03) |
| Noise | general icu | TRUE | 833.60 (136.33) | 512.60 (107.66) | 41.40 (10.19) | 40.60 (11.10) | 0.49 (0.13) | 0.62 (0.09) | 0.07 (0.01) | 0.13 (0.02) | 0.07 (0.01) | 0.58 (0.04) |
| Mask | general icu | TRUE | 717.60 (92.74) | 628.60 (73.13) | 32.20 (2.95) | 49.80 (4.66) | 0.61 (0.04) | 0.53 (0.06) | 0.07 (0.01) | 0.13 (0.01) | 0.07 (0.01) | 0.59 (0.03) |
| | general icu | FALSE | 518.60 (427.31) | 827.60 (400.44) | 25.20 (25.03) | 56.80 (22.97) | 0.70 (0.29) | 0.38 (0.31) | 0.07 (0.01) | 0.12 (0.01) | 0.07 (0.02) | 0.56 (0.07) |
| Reconstract | general icu | TRUE | 532.60 (729.63) | 813.60 (743.10) | 33.20 (45.48) | 48.80 (44.75) | 0.60 (0.55) | 0.40 (0.55) | 0.03 (0.03) | 0.06 (0.06) | 0.06 (0.01) | 0.52 (0.03) |
| | general icu | FALSE | 679.80 (128.29) | 666.40 (119.66) | 31.20 (6.02) | 50.80 (3.83) | 0.62 (0.06) | 0.50 (0.09) | 0.07 (0.01) | 0.13 (0.02) | 0.07 (0.01) | 0.58 (0.03) |

Table 2: Augmentation Performance metrics reported as mean (std).

In Table 2, we investigate how oversampling combined with different augmentation strategies, as well as the use of pretrained weights, affects rare event detection performance. Based on the observations from Table 1, where pretraining on the general ICU cohort yielded more consistent results than pretraining on the trauma cohort, we restrict the comparison here to models trained without pretraining and models pretrained on the general ICU cohort. For augmentation, we include the widely used SMOTE baseline, and for time-series augmentation we evaluate two common strategies: time warping and additive noise. In addition, we consider a task-specific augmentation derived from the Masked Autoencoder, which introduces variability through reconstruction-based masking. All experiments in this table are conducted with 5-fold cross-validation on the trauma-only dataset, with minority class samples oversampled to match the majority class size. Again, the importance of pre-training for rare event detection can be easily observed, while the masking and reconstruction-based augmentation provides better performance by covering more data variance for the minor class.

| Model | tn | fp | fn | tp | sensitivity | specificity | precision | f1_score | PR_auc | ROC_auc |
|---|---|---|---|---|---|---|---|---|---|---|
| 1encoder | 844.20 (129.66) | 819.60 (118.43) | 32.80 (9.23) | 55.20 (8.87) | 0.63 (0.10) | 0.51 (0.08) | 0.06 (0.01) | 0.11 (0.01) | 0.06 (0.01) | 0.59 (0.02) |
| TCNNonly | 617.20 (99.47) | 729.00 (109.88) | 25.40 (5.86) | 56.60 (8.74) | 0.69 (0.08) | 0.46 (0.08) | 0.07 (0.01) | 0.13 (0.02) | 0.08 (0.02) | 0.60 (0.03) |
| GRUonly | 1098.20 (190.05) | 248.00 (214.97) | 62.60 (17.40) | 19.40 (15.19) | 0.24 (0.19) | 0.82 (0.16) | 0.10 (0.06) | 0.10 (0.04) | 0.07 (0.01) | 0.56 (0.05) |

Table 3: Performance of different model architectures reported as mean (std).

In Table 3, we conduct an ablation study to evaluate the contribution of different encoder design choices and the effectiveness of using a single versus a dual encoder. The dual encoder setup leverages both current-night observations and delta values, while the single encoder setup only processes the current-night input stream. All experiments in this table are initialized with pretrained weights from the general ICU cohort to ensure consistent feature representations. For the downstream classification task, we apply random oversampling to balance the classes, but no additional data augmentation is used. This design allows us to isolate the effects of encoder architecture and input configuration on performance. We can observe that after pre-training on general ICU data, the contribution of delta values between two nights is subtle, while the combination of TCNN and GRU as the backbone is important to reduce the bias between the minor and major classes. In general, we can notice great future opportunities to advance using this dataset.

## 5 Conclusions

This work constructs a standardized trauma cohort with reliable and reproducible post-traumatic sepsis onset labels. Comprehensive experiments show promising advances using the new dataset in both post-traumatic sepsis and rare event early detection, which also shows the importance of reliable labels even on a relatively small dataset. In the future, the same pipeline can be extended to other databases (e.g., MIMIC-IV, eICU) to further improve the generalizability of our dataset.

## 6 ETHICS STATEMENT

Our dataset is derived from the publicly available MIMIC-III database, which has been fully de-identified and adheres to HIPAA standards. Access to MIMIC-III requires credentialed approval, and consequently, access to our derived dataset follows the same process. Thereafter, this work has no potential ethical issues to disclose.

## 7 REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide detailed tutorials on how to obtain MIMIC-III access and reproduce our dataset construction step by step. In addition, we supply extensive documentation covering both the clinical rationale behind critical labeling decisions and explanatory comments in the code base. This ensures that the dataset is accessible and interpretable to both machine learning researchers and clinical investigators as shown in the anonymous Github project at `https://github.com/ML4Health-AnonResearch/PostTraumaticSepsis.git`.

## REFERENCES

Roy Adams, Katharine E Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature medicine*, 28(7): 1455–1460, 2022.

Nicholas A Bosch, Anica C Law, Justin M Rucci, Daniel Peterson, and Allan J Walkey. Predictive validity of the sequential organ failure assessment score versus claims-based scores among critically ill patients. *Annals of the American Thoracic Society*, 19(6):1072–1076, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Emanuel Eguia, Corinne Bunn, Sujay Kulshrestha, Talar Markossian, Ramon Durazo-Arvizu, Marshall S Baker, Richard Gonzalez, Faraz Behzadi, Matthew Churpek, Cara Joyce, et al. Trends, cost, and mortality from sepsis after trauma in the united states: an evaluation of the national inpatient sample of hospitalizations, 2012–2016. *Critical care medicine*, 48(9):1296–1303, 2020.

Hadj Ali Elmerahi, Baghdad Atmani, Fatiha Barigou, Belarbi Khemliche, Badreddine Errouane, and Mohammed Bousmaha. Parrallel lstm-dnn fusion model for early prediction of sepsis in intensive care units. In *2024 4th International Conference on Embedded Distributed Systems (EDiS)*, pp. 43–48, 2024. doi: 10.1109/EDiS63605.2024.10783411.

Jesper Eriksson, Mikael Eriksson, Olof Brattström, Elisabeth Hellgren, Ola Friman, Andreas Gidlöf, Emma Larsson, and Anders Oldner. Comparison of the sepsis-2 and sepsis-3 definitions in severely injured trauma patients. *Journal of Critical Care*, 54:125–129, 2019.

Kevin Ewig, Xiangwen Lin, Tucker Stewart, Katherine Stern, Grant O'Keefe, Ankur Teredesai, and Juhua Hu. Multi-subset approach to early sepsis prediction. In *2023 Congress in Computer Science, Computer Engineering, Applied Computing (CSCE)*, pp. 1335–1341, 2023a. doi: 10.1109/CSCE60160.2023.00224.

Kevin Ewig, Xiangwen Lin, Tucker Stewart, Katherine Stern, Grant O'Keefe, Ankur Teredesai, and Juhua Hu. Multi-subset approach to early sepsis prediction. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, pp. 1335–1341. IEEE, 2023b.

Carolin Fleischmann, André Scherag, Neill KJ Adhikari, Christiane S Hartog, Thomas Tsaganos, Peter Schlattmann, Derek C Angus, and Konrad Reinhart. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *American journal of respiratory and critical care medicine*, 193(3):259–272, 2016.

MIT Laboratory for Computational Physiology. Mimic code repository. `https://github.com/MIT-LCP/mimic-code`, 2025. Accessed: 2025-09-03.

Mengsha Fu, Jiabin Yuan, and Chen Bei. Early sepsis prediction in icu trauma patients with using an improved cascade deep forest model. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 634–637. IEEE, 2019.

Faheem W Guirgis, Scott Brakenridge, Selina Sutchu, Jay D Khadpe, Taylor Robinson, Richard Westenbarger, Stephen T Topp, Colleen J Kalynych, Jennifer Reynolds, Sunita Dodani, et al. The long-term burden of severe sepsis and septic shock: Sepsis recidivism and organ dysfunction. *Journal of Trauma and Acute Care Surgery*, 81 (3):525–532, 2016.

Kucun Guo, Bao Pan, Xinliang Zhang, Dezheng Hu, Guangyue Xu, Lin Wang, and Shimin Dong. Developing an early warning system for detecting sepsis in patients with trauma. *International Wound Journal*, 21(1):e14652, 2024.

Dara L Horn, Michael Mindrinos, Kirsten Anderson, Sujatha Krishnakumar, Chunlin Wang, Ming Li, Jill Hollenbach, and Grant E O'Keefe. Hla-a locus is associated with sepsis and septic shock after traumatic injury. *Annals of surgery*, 275(1):203–207, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Jerome Aboab, Jesse D Raffa, Tom J Pollard, Rodrigo O Deliberato, Leo A Celi, and David J Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine*, 46 (4):494–499, 2018.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine*, 104:101820, 2020a.

Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature communications*, 11(1):3852, 2020b.

Seunghee Lee, Geonchul Shin, Jeongseok Hwang, Yunjeong Hwang, Hyunwoo Jang, Ju Han Park, Sunmi Han, Kyeongmin Ryu, and Jong-Yeup Kim. Early prediction of sepsis in the intensive care unit using the gru-d-mgp-tcn model. *IEEE Access*, 12:148294–148304, 2024. doi: 10.1109/ACCESS.2024.3470851.

Jiang Li, Fengchan Xi, Wenkui Yu, Chuanrui Sun, Xiling Wang, et al. Real-time prediction of sepsis in critical trauma patients: Machine learning–based modeling study. *JMIR Formative Research*, 7(1):e42452, 2023.

Xiang Li, Yanni Kang, Xiaoyu Jia, Junmei Wang, and Guotong Xie. Tasp: A time-phased model for sepsis prediction. In *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4, 2019. doi: 10.22489/CinC.2019.049.

Luchen Liu, Haoxian Wu, Zichang Wang, Zequn Liu, and Ming Zhang. Early prediction of sepsis from clinical data via heterogeneous event aggregation. In *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4, 2019. doi: 10.22489/CinC.2019.157.

Joseph P Minei, Joseph Cuschieri, Jason Sperry, Ernest E Moore, Michael A West, Brian G Harbrecht, Grant E O'Keefe, Mitchell J Cohen, Lyle L Moldawer, Ronald G Tompkins, et al. The changing pattern and implications of multiple organ failure after blunt injury with hemorrhagic shock. *Critical care medicine*, 40(4):1129–1135, 2012.

James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, and Terry Lyons. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. In *2019 Computing in Cardiology (CinC)*, pp. Page 1–Page 4, 2019. doi: 10.22489/CinC.2019.014.

Naoki Nonaka and Jun Seita. Demographic information initialized stacked gated recurrent unit for an early prediction of sepsis. In *2019 Computing in Cardiology (CinC)*, pp. 1–4, 2019. doi: 10.22489/CinC.2019.153.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo Anthony Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5: 180178, 2018.

Guilherme Ramos, Erida Gjini, Luis Coelho, and Margarida Silveira. Unsupervised learning approach for predicting sepsis onset in icu patients. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 1916–1919, 2021. doi: 10.1109/EMBC46164.2021.9629559.

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.

Chanu Rhee and Michael Klompas. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(Suppl 1):S89, 2020.

Chanu Rhee, Maximilian S Jentzsch, Sameer S Kadri, Christopher W Seymour, Derek C Angus, David J Murphy, Greg S Martin, Raymund B Dantes, Lauren Epstein, Anthony E Fiore, et al. Variation in identifying sepsis and organ dysfunction using administrative versus electronic clinical data and impact on hospital outcome comparisons. *Critical care medicine*, 47(4):493–500, 2019.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219): 200–211, 2020.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Katherine Stern, Qian Qiu, Michael Weykamp, Grant O'Keefe, and Scott C Brakenridge. Defining posttraumatic sepsis for population-level research. *JAMA Network Open*, 6(1):e2251445–e2251445, 2023.

Tucker Stewart, Katherine Stern, Grant O'Keefe, Ankur Teredesai, and Juhua Hu. Nprl: Nightly profile representation learning for early sepsis onset prediction in icu trauma patients. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 1843–1852. IEEE, 2023.

Ankur Teredesai, Sijin Huang, Tucker Stewart, Juhua Hu, Armaan Thakker, Katherine Stern, and Grant E O'Keefe. Sub-sequence graph representation learning on high variability data for dynamic risk prediction in critical care. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 2082–2092. IEEE, 2022.

Luan Tran, Manh Nguyen, and Cyrus Shahabi. Representation learning for early sepsis prediction. In *2019 Computing in Cardiology (CinC)*, pp. 1–4, 2019. doi: 10.22489/CinC.2019.021.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Sijia Zhang, Zunliang Wang, Yuyan Zhang, and Songqiao Liu. A reinforcement learning approach for predicting the onset of septic shock patients with unfair bias. In *2024 8th International Conference on Biomedical Engineering and Applications (ICBEA)*, pp. 182–187, 2024. doi: 10.1109/ICBEA62825.2024.00041.

# A Appendix

## A LLM usage Statement

We did not use LLM at all during the idea stage of this work, while LLM was used to detect general writing problems such as typos, grammar errors, and inappropriate phrasing.

## B Cohort Details

### B.1 ICD9 Code for Trauma

To extract a reliable patient cohort of critically ill patients with trauma from the MIMIC-III dataset, we use a well-defined list of ICD9-Ecode, that is, E8000-E8480, E8800-E9057, E9060-9259, E9270-E9289, E9507, E9520, E9521, E9520-E9589, E9600, E9610- E9689, E9680-E9689, E9700-E9760, E9780-E9799, E9806, E9830-E9886, E9888, E9889, E9900-E9961, and E9968-E9989.

We explicitly exclude poisoning-related codes (e.g., E850–E854) to ensure compatibility with trauma-specific datasets that may be developed in future studies. These codes are also excluded because patients with poisoning are less likely to develop sepsis, which is the focus of our early detection task.

### B.2 Sepsis Ratio and Ventilation Days

Fig. 2 illustrates the relationship between the sepsis ratio and the number of ventilation days in the MIMIC-III dataset. The x-axis represents the total number of days a patient received mechanical ventilation during a single admission, while the y-axis denotes the corresponding sepsis ratio. The figure highlights a rapid increase in the sepsis ratio starting from patients with three or more days of mechanical ventilation, supporting the rationale for using this threshold to identify critically ill trauma patients at higher risk of developing sepsis.

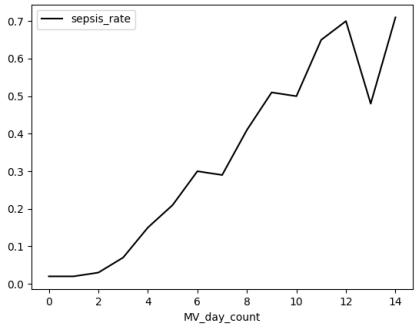


Figure 2: Sepsis ratio vs. number of ventilation days.
A chart illustrating the relationship between sepsis ratio and the number of ventilation days. The x-axis represents ventilation days, while the y-axis denotes the corresponding sepsis ratio. The sepsis ratio shows a rapid increase starting from patients with three days of ventilation.

### B.3 Cohort Selection Flow Diagram

Fig. 3 presents the step-by-step inclusion and exclusion criteria used to refine the trauma cohort for this study. The diagram outlines the sequential filtering process from the initial dataset, highlighting key inclusion and exclusion steps that resulted in the final cohort of 1,570 admissions(HADM_ID).

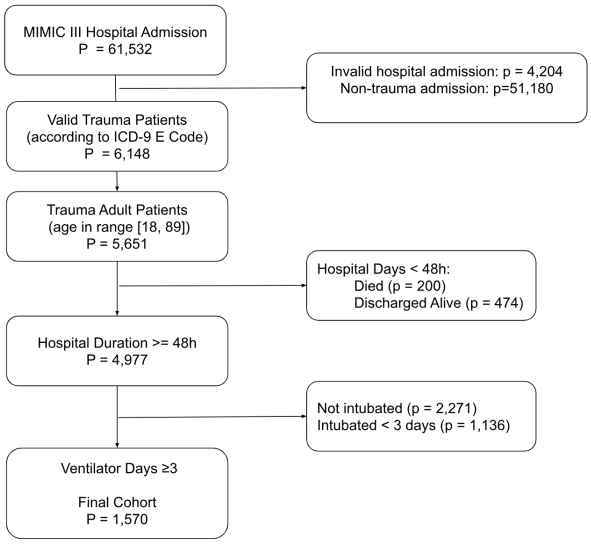


Figure 3: Flow diagram of study inclusion and exclusion criteria.
A flow diagram illustrating the study's inclusion and exclusion criteria. The diagram outlines the sequential filtering of patients from the initial dataset, detailing key inclusion criteria and exclusion steps leading to the final study cohort.

### B.4 PATIENT CHARACTERISTICS

Table 4: Characteristics and Clinical Outcomes of Critically Ill Injured Patients Admitted to Intensive Care Unit (2012-2020)

| **Characteristic** | **No. (%) (N = 1570[a])** |
|---|---|
| Age, median (IQR), y | 59 (43-75) |
| Sex, n (%) | |
| Female | 538 (34) |
| Male | 1032 (66) |
| Race, n (%) | |
| Asian | 22 (1.4) |
| Black | 85 (5.4) |
| White | 1102 (70.2) |
| Other | 75 (4.8) |
| Hispanic | 63 (4.0) |
| Unknown | 223 (14.2) |
| Charlson Comorbidity Index[c], median (IQR) | 12 (0-12) |
| Injury Severity Score, median (IQR) | 16 (4-20) |
| Body regions with an AIS $\geq 3$, n(%) | |
| Head/Neck | 709 (45) |
| Chest | 395 (25) |
| Abdomen | 135 (9) |
| Lower extremity | 215 (14) |
| LACTATE mmol/L[d], median (IQR) | 3 (2-5) |
| Unknown, n (%) | 211 (13) |
| Initial ED SBP $< 90$ mm Hg, n (%) | 72 (5) |
| **Outcomes** | |
| Sum of ICU days, median (IQR) | 10 (6-18) |
| Hospital admission days, median (IQR) | 16 (10-24) |
| Mechanical ventilation days, median (IQR) | 7 (4-12) |
| Discharged Location, n (%) | |
| Rehab/Distinct | 588 (37) |
| Dead | 341 (22) |
| Home | 289 (18) |
| SNF | 279 (18) |
| Other | 73 (5) |

**Abbreviations:**
AIS: Abbreviated Injury Severity Score
ICU: Intensive care unit
LTFC: Long-term care facility
SBP: Systolic blood pressure
SNF: Skilled nursing facility
ED: Emergency Department
[a] Missing values greater than 5% are reported.
[c] Elixhauser-van Walraven comorbidity index.
[d] Highest value documented during the first 48 hours of ED admission.

After necessary refinement as mentioned in the main paper, we obtained a total of 1,570 admissions as the cohort of critically ill trauma patients suitable for early sepsis onset detection. The detailed characteristics for these patients are summarized in Table 4. We can observe that of the 1,570 admissions that met the inclusion criteria, the median age for the patients was 59 years (IQR, 43-75 years). Ethnicity distribution was as follows: 70.2% White, 14.2% Unknown, 5.4% Black, 4.8% Other, 4.0% Hispanic, and 1.4% Asian. Gender distribution was 66% male and 34% female. Moreover, the median length of hospital stay was 16 days (IQR, 10-24 days), the median length of ICU stay was 10 days (IQR, 6-18 days), and the median number of days on ventilation was 7 days (IQR, 4-12 days). Finally, hospital mortality data indicated that 22% (341 individuals) died during hospitalization, while 78% (1,229 individuals) survived to discharge. Among those discharged, 18% (279 individuals) were sent to skilled nursing facilities or long-term care facilities, and 18% (289 individuals) were discharged to home without assistance.

# C Sepsis Identification Details

## C.1 Antibiotic List

This appendix provides a complete list of antibiotics used in the study. The antibiotics were selected based on established guidelines and prior research (Stern et al., 2023; Johnson et al., 2016; Komorowski et al., 2018), ensuring relevance to trauma-related sepsis. Table 5 detail the included antibiotics, their administration routes, and whether they are classified as prophylactic.

Table 5: List of Drugs

| gsn | drug | route | isProphylactic |
|---|---|---|---|
| 8854 | Penicillin G Potassium | IV | 0 |
| 8920 | Ampicillin-Sulbactam | IV | 1 |
| 8921 | Ampicillin-Sulbactam | IV | 1 |
| 8921 | Ampicillin-Sulbactam | IV | 1 |
| 8932 | Ampicillin Sodium | IV | 0 |
| 8935 | Ampicillin Sodium | IV | 0 |
| 8937 | Ampicillin Sodium | IV | 0 |
| 8965 | Nafcillin | IV | 0 |
| 9143 | Cefuroxime Sodium | IV | 0 |
| 9144 | Cefuroxime Sodium | IV | 0 |
| 9156 | CefTRIAXone | IV | 0 |
| 9156 | Ceftriaxone | IV | 0 |
| 9157 | CefTRIAXone | IV | 0 |
| 9157 | Ceftriaxone | IV | 0 |
| 9162 | CeftriaXONE | IV | 0 |
| 9165 | CeftriaXONE | IV | 0 |
| 9171 | CefTAZidime | IV | 0 |
| 9172 | CefTAZidime | IV | 0 |
| 9172 | CeftazIDIME | IV | 0 |
| 9181 | Cefotetan | IV | 0 |
| 9181 | Cefotetan | IV | 0 |
| 9221 | Doxycycline Hyclate | IV | 0 |
| 9251 | Erythromycin | IV | 1 |
| 9251 | Erythromycin Lactobionate | IV | 1 |
| 9252 | Erythromycin | IV | 1 |
| 9289 | Gentamicin | IV | 0 |
| 9291 | Gentamicin | IV | 0 |
| 9294 | Gentamicin | IV | 0 |
| 9299 | Gentamicin | IV | 0 |
| 9299 | Gentamicin Sulfate | IV | 0 |
| 9312 | Amikacin | IV | 0 |
| 9328 | Vancomycin | IV | 0 |
| 9329 | Vancomycin | IV | 0 |
| 9329 | Vancomycin HCl | IV | 0 |
| 9329 | Vancomycin Oral Liquid | PO | 0 |
| 9329 | Vancomycin Oral Liquid | PO/NG | 0 |
| 9330 | Vancomycin | IV | 0 |
| 9331 | Vancomycin | IV | 0 |
| 9331 | Vancomycin Enema | IV | 0 |
| 9331 | Vancomycin HCl | IV | 0 |
| 9344 | Clindamycin | IV | 0 |
| 9344 | Clindamycin Phosphate | IV | 0 |
| 9361 | Aztreonam | IV | 0 |
| 9362 | Aztreonam | IV | 0 |
| 9365 | Imipenem-Cilastatin | IV | 0 |

Continued on next page

Table 5: List of Drugs

| gsn | drug | route | isProphylactic |
|---|---|---|---|
| 9393 | Sulfameth/Trimethoprim | IV | 0 |
| 9393 | Sulfamethoxazole-Trimethoprim | IV | 0 |
| 9525 | Amphotericin B | IV | 0 |
| 9588 | MetRONIDAZOLE (FLagyl) | IV | 0 |
| 9588 | Metronidazole | IV | 0 |
| 9592 | Metronidazole | IV | 0 |
| 13052 | Clindamycin | IV | 0 |
| 13053 | Clindamycin | IV | 0 |
| 13645 | Rifampin | IV | 0 |
| 14196 | Aztreonam | IV | 0 |
| 14197 | Aztreonam | IV | 0 |
| 15327 | Gentamicin | IV | 0 |
| 15355 | Nafcillin | IV | 0 |
| 15538 | CeftAZIDime | IV | 0 |
| 15538 | CeftazIDIME | IV | 0 |
| 15538 | Ceftazidime | IV | 0 |
| 15539 | CeftAZIDime | IV | 0 |
| 15539 | CeftazIDIME | IV | 0 |
| 15539 | Ceftazidime | IV | 0 |
| 15920 | Ciprofloxacin IV | IV | 0 |
| 15921 | Ciprofloxacin | IV | 0 |
| 15921 | Ciprofloxacin IV | IV | 0 |
| 15932 | Penicillin G Potassium | IV | 0 |
| 15933 | Penicillin G Potassium | IV | 0 |
| 15934 | Penicillin G Potassium | IV | 0 |
| 21185 | Piperacillin-Tazobactam | IV | 0 |
| 21185 | Piperacillin-Tazobactam Na | IV | 0 |
| 21187 | Piperacillin-Tazobactam | IV | 0 |
| 21187 | Piperacillin-Tazobactam Na | IV | 0 |
| 21701 | Cefotetan | IV | 0 |
| 21702 | Cefotetan | IV | 0 |
| 24094 | CefePIME | IV | 0 |
| 24095 | CefePIME | IV | 0 |
| 26488 | Meropenem | IV | 0 |
| 26489 | Meropenem | IV | 0 |
| 27468 | CefePIME | IV | 0 |
| 27468 | Cefepime | IV | 0 |
| 27470 | Cefepime | IV | 0 |
| 29925 | Levofloxacin | IV | 0 |
| 29927 | Levofloxacin | IV | 0 |
| 29928 | Levofloxacin | IV | 0 |
| 29929 | Levofloxacin | IV | 0 |
| 31452 | Azithromycin | IV | 0 |
| 31535 | Nafcillin | IV | 0 |
| 40819 | Piperacillin-Tazobactam | IV | 0 |
| 40819 | Piperacillin-Tazobactam | IV | 0 |
| 40819 | Piperacillin-Tazobactam Na | IV | 0 |
| 40819 | Piperacillin-Tazobactam Na | IV | 0 |
| 43952 | Vancomycin | IV | 0 |
| 43952 | Vancomycin HCl | IV | 0 |
| 45131 | Linezolid | PO/NG | 0 |
| 45131 | Linezolid | PO | 0 |
| 45134 | Linezolid | IV | 0 |
| 45134 | Linezolid | PO/NG | 0 |
| 46770 | Levofloxacin | IV | 0 |

Continued on next page

Table 5: List of Drugs

| gsn | drug | route | isProphylactic |
|---|---|---|---|
| 46771 | Levofloxacin | IV | 0 |
| 57824 | Ciprofloxacin | IV | 0 |
| 57825 | Ciprofloxacin | IV | 0 |
| 57825 | Ciprofloxacin IV | IV | 0 |
| 59424 | CefTRIAXone | IV | 0 |
| 59424 | CeftriaXONE | IV | 0 |
| 59425 | CefTRIAXone | IV | 0 |
| 59425 | CeftriaXONE | IV | 0 |
| 59747 | CefTAZidime | IV | 0 |
| 59747 | CeftazIDIME | IV | 0 |

### C.2 Sepsis Onset Timing Distribution

Fig. 4 illustrates the distribution of sepsis onset timing relative to hospital admission among the identified 535 sepsis cases. The x-axis represents time in days, while the y-axis indicates the number of patients experiencing sepsis onset on each day. Notably, the peak for culture orders, marking the onset of sepsis, occurs on day 5, aligning with clinical observations (Horn et al., 2022). This histogram provides insight into the temporal pattern of sepsis onset in the trauma cohort.

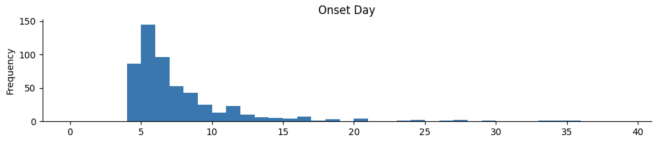


Figure 4: Timing of sepsis.
A chart showing the distribution of sepsis onset timing relative to hospital admission. The x-axis represents time in days, while the y-axis indicates the number of patients. Notably, the peak for culture orders, indicating the onset of sepsis, occurs on day 5

### C.3 Timeline visualization of post-trauma sepsis onset assignment criteria.

Figure 5 illustrates the timeline-based criteria for labeling hospital-acquired post-trauma sepsis. The sepsis onset timestamp (date + time) is defined as the order time of the earliest blood culture that satisfies all of the following criteria: 1. Hospital-Acquired Sepsis: The sepsis onset timestamp comes from the culture order timestamp (date + time). To qualify, the culture must occur at least 72 hours (3 days) after hospital admission. 2. Suspicion of Infection: A new antibiotic must be ordered within a 5-day window, defined as 2 days before to 2 days after the earliest blood culture is ordered. The antibiotic must then be administered for at least 4 consecutive days, or until the patient is discharged or deceased. 3. Organ Failure: There must be an increase of at least 2 points in the modified SOFA score within a 7-day window, defined as 3 days before to 3 days after the culture is ordered.

### C.4 Temporal Trends of Physiological Features Before Sepsis Onset

To visualize early signs of sepsis, we plot the temporal trends of key physiological features (e.g., heart rate, blood pressure) in the four days leading up to sepsis onset. The x-axis represents time relative to sepsis onset, measured in Accumulated Days Before Sepsis Onset (DBSO), while the y-axis shows the average physiological measurements across all sepsis patients. In this representation, x = -4 corresponds to four days before sepsis onset, and x = 0 marks the onset of sepsis. The range from x = -1 to 0 represents positive nighttime samples.

As shown in Figure 6, certain features, particularly temperature (Temp) and heart rate (HR), exhibit noticeable changes at the beginning of the positive night (x = -1). Based on this observation and individual patient physiological trends, we hypothesize that the delta values—the difference between a patient's current night and their previous night—can provide valuable information for early sepsis detection. Therefore, we compute these delta values as an auxiliary input to enhance the model's ability to capture early signs of sepsis.

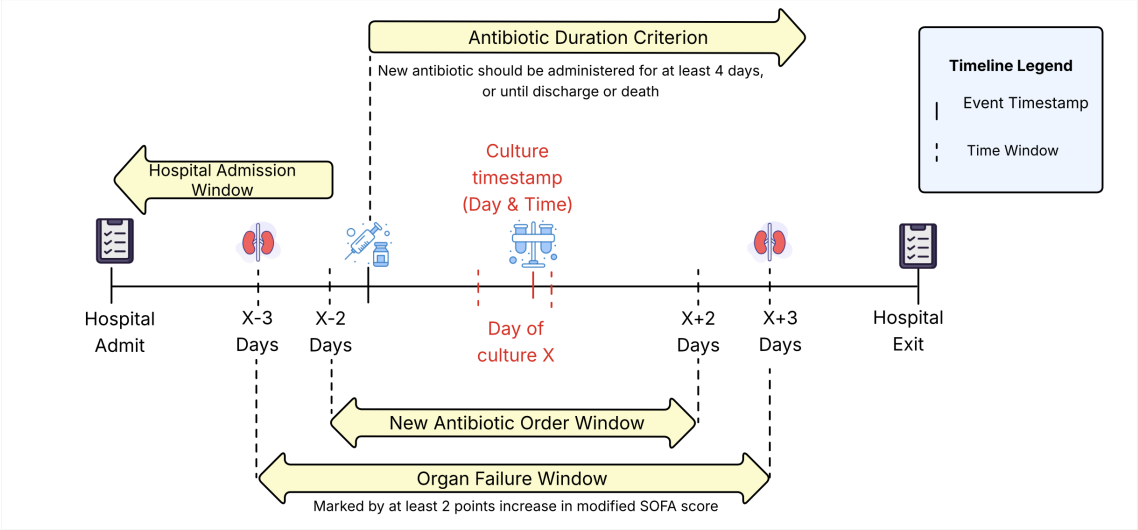


Figure 5: Timeline visualization of post-trauma sepsis onset assignment criteria.

Notably, since sepsis onset can occur at different times for different patients, the amount of available data decreases as we look further back in time. For instance, we have sufficient patient data to compute meaningful averages one day before sepsis onset, but the number of patients with data available ten days before onset is significantly smaller. To ensure reliable visualization, we limit the analysis to the four days leading up to sepsis onset.

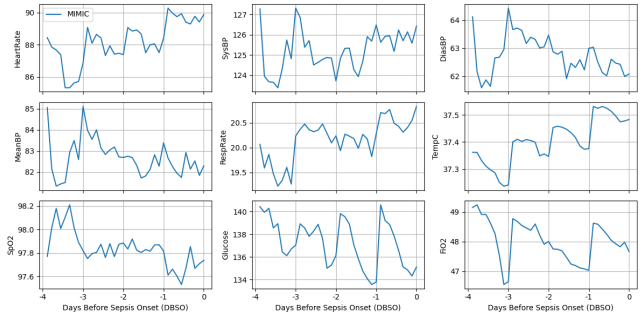


Figure 6: Temporal trends of key physiological features in the four days leading up to sepsis onset. The x-axis represents time relative to sepsis onset, measured in Accumulated Days Before Sepsis Onset (DBSO), while the y-axis shows the average physiological measurements across all patients.

A set of line plots visualizing physiological trends before sepsis onset. Each subplot represents a different feature, including heart rate, blood pressure, respiratory rate, temperature, SpO2, glucose, and FiO2. The x-axis shows time in days before sepsis onset, with x = 0 marking sepsis onset. Notable changes in heart rate and temperature are observed around x = -1.

## D Detection Setup

Since handling irregular time-series data and imputing missing values are widely discussed topics in time-series analysis, we also include the ”N dataset” (which retains NaN values) to preserve the rawest form of the data without introducing potential imputation bias. Notably, the ”S dataset” is a subset of the ”N dataset” since some instances cannot be fully filled using our method and window constraints. Additionally, if no recorded data exists for a specific patient during a given night in the ”N dataset,” that instance cannot appear in the ”S dataset” either. This ensures that we do not introduce artificially generated values that would be identical across all timestamps, which could mislead the model. In the N dataset, we excluded several instances for specific reasons: 9 patients had data limited to nightly observations after day 14 (6 of whom were sepsis patients), which falls outside the critical period for early sepsis detection. Additionally, we removed 26 sepsis patients who only had data available after their sepsis onset and 48 pa-

tients for whom we could not locate positive instances. These positive instances were often located outside ICU stays, occurring before, between, or after ICU stays. Since vital signs in the MIMIC dataset are only documented during ICU stays, this complicates the tracking of early sepsis signs (i.e. the positive instance). Furthermore, excluding data from outside the ICU is justified, as these patients may have transitioned between units or been hospitalized for extended periods prior to ICU admission, potentially altering the reasons for their sepsis. Ultimately, our final dataset for the N subset includes 455 positive instances and 8,522 negative instances, totaling 8,977 cases across 1,535 unique patients.

### D.1 Cross-Validation Data Splitting Strategy

The data splitting process aims to prevent data leakage and ensure a balanced representation of sepsis and non-sepsis patients across subsets. In the MIMIC dataset, each individual is assigned a unique SUBJECT_ID, which may encompass one or multiple hospital admissions. The time gaps between these admissions can vary significantly, as patients may be admitted for different reasons. Although the hospital admission ID (HADM_ID) aligns more closely with the clinical concept of a patient, we perform the split based on SUBJECT_ID. Specifically, we define a sepsis individual as one for whom at least one hospital admission is identified as sepsis, and we then execute a stratified split based on SUBJECT_ID to ensure that all records for each individual remain within the same subset.

Given the limited data size, we employ stratified 5-fold cross-validation to ensure balanced label distribution across folds. In each iteration, one fold is used for testing, while the remaining folds are used for training. Since the dataset is small, the same fold serves both as the validation set during training and as the test set for final evaluation. The key difference is that oversampling is applied to the validation set during training to address class imbalance, whereas the test set remains untouched to preserve the original data distribution. Across all folds, the class imbalance ratio (positive/total) ranges around 0.05. For further details, refer to Table 6 for "S Dataset" and Table 7 for "N Dataset".

| Fold | Total Instances | Positive Instances | Negative Instances | Imbalance Ratio |
|---|---|---|---|---|
| 0 | 1735 | 88 | 1647 | 0.0507 |
| 1 | 1763 | 87 | 1676 | 0.0493 |
| 2 | 1797 | 93 | 1704 | 0.0518 |
| 3 | 1711 | 90 | 1621 | 0.0526 |
| 4 | 1753 | 82 | 1671 | 0.0468 |
| Total | 8759 | 440 | 8319 | 0.0502 |

Table 6: Distribution of the S Dataset across folds.

| Fold | Total Instances | Positive Instances | Negative Instances | Imbalance Ratio |
|---|---|---|---|---|
| 0 | 1788 | 92 | 1696 | 0.0515 |
| 1 | 1813 | 90 | 1723 | 0.0496 |
| 2 | 1825 | 94 | 1731 | 0.0515 |
| 3 | 1740 | 92 | 1648 | 0.0529 |
| 4 | 1811 | 87 | 1724 | 0.0480 |
| Total | 8977 | 455 | 8522 | 0.0507 |

Table 7: Distribution of the N Dataset across folds.