Temporal Modeling of Traumatic Patient for Early Sepsis Onset Prediction as Rare Event in ICU

Tucker Reed Stewart

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Juhua Hu

Ankur Teredesai

Katherine Stern

Program Authorized to Offer Degree:

Computer Science and Systems

University of Washington

Abstract

Temporal Modeling of Traumatic Patient for Early Sepsis Onset Prediction as Rare Event in ICU

Tucker Reed Stewart

Chair of the Supervisory Committee:
Juhua Hu
School of Engineering and Technology

Sepsis is a hyper-inflammatory syndrome that develops in a patient's body in response to the presence of infection. It leads to severe organ dysfunction and is one of the leading causes of mortality that occur in Intensive Care Units (ICUs) worldwide. These complications can be reduced through early application of antibiotics, hence the ability to anticipate the onset of sepsis early is crucial to the survival and well-being of patients. Current machine learning algorithms deployed inside medical infrastructures have demonstrated poor performance and are insufficient for anticipating sepsis onset. In recent years, there has been a substantial effort developing various deep learning methodologies for predicting sepsis, but many fail to capture the time of onset and in a manner suitable for integration into medical facilities. In this study, we propose a temporal deep learning framework that can capture the temporal progressing pattern and predict whether sepsis onset will occur within a 24-hour window using data collected at night, when patient-provider ratios are higher due to cross-coverage resulting in limited observation to each

patient. Moreover, we design a pre-training technique to alleviate the rare event problem of sepsis onset (i.e., the class imbalance problem). Our empirical study using data from a level 1 trauma center demonstrates the effectiveness of our proposed method.

1 Introduction

Sepsis is a syndrome of a host's dysregulated immune response to the presence of infection. The host's immune system is so over zealous in its mobilization against pathogens in the body, that it starts to attack the host's own tissue, causing organ dysfunction and tissue damage. To this day, sepsis remains to be a prominent complication in the modern medical facilities, particularly Intensive Care Units (ICUs). According to a global audit conducted in 2018 [24], depending on the region studied, approximately 13.6% to 39.3% of patients admitted in ICU are affected by sepsis. Globally, this proportion is 29.5%. Of the patients that develop sepsis in the ICU, 25.8% will die likely due to sepsis. From these stats, it can be deduced that approximately 7.6% of patients admitted into ICUs across the globe will die as a result of sepsis, making it one of the leading causes of death in ICUs.

These outcomes can be improved through early intervention [19]. Preemptive administration of antibiotics treats the underlying infection, preventing the sepsis from occurring or at least reducing the symptoms as the presence of the infection starts to decline. It is believed that as much as 80% of sepsis-related deaths are preventable with early intervention and chances of survival decrease about 8% each hour that action is not taken [15]. Hence, it is critical to the survival and well-being of patients to anticipate the onset of sepsis accurately and early. Moreover, in critically ill trauma patients (i.e., individuals admitted to the ICU for management of injury caused by blunt or penetrating force), injury-related inflammation and organ dysfunction may increase the risk for sepsis, while also masking the clinical signs of infection [5, 6]. Thus, detecting sepsis early in the critically ill trauma population is in great need but challenging, which is the application focus of this work.

Due to challenges in identifying septic patients prior to onset of sepsis and the over abundance of available data, many have looked into machine learning using patients' Electronic Health Records (EHRs) to predict sepsis or septic shock [7, 11–13, 16, 17, 19, 22, 23, 26]. EHR, the most commonly used data format in healthcare, records patient data from a variety of different categories and modalities such as patient demographics, comorbidities, indications for hospital admission, lab results, hourly physiology, and therapeutics administered.

Many prior studies work only retrospectively [13]. Specifically, these studies [12, 13, 17, 19] identify the timestamp of the target event such as sepsis onset or septic shock, then look back for a fixed time interval for early prediction. As such, the time length between prediction and the target event is fixed. This analytic approach is less useful in the prospective setting. It is because that we often do not know in advance when we would expect sepsis, which makes it difficult to decide when to use the model for prediction. Thereafter, we may need to use the model very frequently (e.g., every hour), so that we do not miss any chance of early prediction. However, this is not realistic in a live setting. To ensure the viability in live clinical settings, the problem must be framed from the point of prediction, not onset since it is unknown. For this reason, we use data from each night the patient is in the ICU and assess the potential for sepsis each day.

Besides this, these works [12,13,17,19] create one sample per hospital admission, not multiple such as for each day the patient is in the hospital. In this case, the model does not discriminate between days where sepsis is or is not present in the patient. Thus, they do not capture the time of onset.

Beyond the design of the methodological framing, there is the methodology for modeling EHR data for predicting medical outcomes like sepsis onset. Based on the methodological framing employed, patients' EHRs from a prescribed window of time are given as input for prediction. Some studies have opted to use traditional machine learning methods such as Random Forest [12] and XGBoost [19] to predict whether sepsis will onset n hour from when the last observation of input data was recorded. However, traditional modeling techniques like these fail to consider the time or sequencing features are ordered in. As such, most other researches utilize temporally conscious deep learning architectures such as Recurrent Neural Networks (RNNs) [11,19,23,26] and Temporal Convolutional Neural Networks (TCNs) [17,22] which can capture the temporal and sequential attributes of the data as well. RNNs are ideal for sequential data such as the temporal EHR data we are using this study. This form of deep learning architectures utilize recurrent units to convey relevant information from past observations to the current which captures the temporal relation between samples. In comparative studies [19], RNNs significantly outperformed traditional machine learning methods due to their ability to integrate the temporal elements of EHR for prediction; motivating the need for temporal modeling techniques in early sepsis onset prediction.

A prominent challenge associated with the early prediction of sepsis onset is the rare event problem which has not been sufficiently addressed in the literature for early sepsis onset prediction. When applying temporal modeling techniques prospectively, sepsis onset becomes a rare event. With the occurrence rate of sepsis being between 13.6% and 39.3% in the ICU, this may not seem like a rare event but this is at the resolution of the visits. To capture the time of sepsis onset within the visit, we must increase the granularity of the prediction task. Instead of labeling the entire visit as sepsis or no sepsis, each day within the visit is examined for the first occurrence of sepsis onset. Time prior to sepsis onset is treated as time at risk, which is appropriate. However, in this format the number of examples available to learn are very few compared to the number of negative examples creating a serious class imbalance. This class imbalance is problematic for modeling because it leads the machine learning algorithm to prioritize the majority class in order to minimize the empirical error.

In this study, we look to 1) develop a methodology that is applicable to clinical settings and can be later deployed in hospitals to aid in the detection of sepsis onset within 24 hours and 2) address challenges with predicting sepsis onset as a rare event. We propose a Multi-Modal RNN that uses data collected at night, between the hours of 10 p.m. and 6 a.m. when observation to each patient is limited, to predict the occurrence of sepsis onset within the next 24 hours. Deep learning are exceptional in their ability to model non-linear patterns in data and RNNs are able to capture the temporal information from sequential

EHR data. Using a Multi-Modal RNN, we are able to model the temporality of hourly physiology as well as incorporate static features to help inform the model of potential risk factors. Then to solve data imbalance that results from sepsis onset being a rare event, we pre-train the Multi-Modal RNN which has been demonstrated to help train deep learning models with imbalanced data [2,9,14]. Our empirical study using de-identified traumatic patients from year 2012 to 2019 at a level 1 trauma center demonstrates the effectiveness of our proposed method as a pre-screening assistant each morning.

2 Related Work

In this section, we briefly review different sepsis prediction setups used in the literature and different machine learning methodologies used for early sepsis prediction.

2.1 Historical Sepsis Prediction Setup

In the literature, most studies used a 'fixed time to onset' framing for their prediction setup [12, 13, 17, 19]. For example, Khojandi et al. [12] used the first 12, 24, and 48 hours of data after admission to predict whether the patient will develop sepsis at any time during their stay. It is not a surprise that the variability in time lags between the observation window and time of onset made the performance very weak. As a result, the authors also tried predicting onset using data 12, 24, and 48 hours prior to the occurrence of onset which achieved much higher results and were commonly used. However, as mentioned by Liu et al. [19], this setup is learning to predict onset at fixed intervals, and thus they are often limited to these retrospective studies where the time of onset is known a priori. Similarly, some other works used the entire observation data from admission up to the point of prediction [7, 16, 22, 26] which Lauritsen et al. [18] denotes as sequential framing. However, both of these framings require that the prediction happens very frequently, so that we do not miss any chance to do early prediction. More importantly, they did not consider the real needs in medical infrastructures. At ICU, it is often during the night time when the patient-provider ratio are much higher due to cross-coverage, we would have very limited observation to each patient. Therefore, an assistant prediction from machine learning every morning before we have full observation of each patient is greatly valuable. In this work, we focus on whether sepsis onset will occur within a 24-hour window using data collected at night, which can be utilized every morning and more practical.

2.2 Machine Learning for Early Sepsis Prediction

Traditional machine learning methods have been applied for sepsis onset prediction, like Random Forest [12] and XGBoost [19]. The most famous model deployed in hospitals is Epic Sepsis Model (ESM), which is a penalized logistic

regression model that produces a score indicating the potential for sepsis which is calculated every fifteen minutes. Recently however, Wong et al. [32] found that ESM performed horrendously using data from Michigan Medicine collected between December 2018 to October 2019. While there is still some utility of ESM scores to help medical staffs' decisions, which is why is was adopted, there is a need for better early predictions of sepsis onset to be deployed in medical infrastructures. Moreover, these traditional machine learning algorithms often require extensive manual effort on the part of feature engineering to get the best results.

Thereafter, due to the development of deep neural networks and their advances in automatic feature learning, more and more deep learning methods have been applied to sepsis related prediction. For example, Liu et al. [19] compared Generalized Linear Model (GLM), XGBoost, and RNN in predicting septic shock with data up to 12 hours prior to shock, where RNN shows the best performance by a significant margin. Scherpf et al. [26] also utilized an RNN for predicting sepsis onset with data three hours prior to the onset of sepsis. However, deep learning architectures are often difficult to interpret. One way of overcoming this is to use an attention mechanism with RNN, which has been adopted by Kaji et al. [11]. The attention mechanism helps to improve performance while also adding some level of interpretability for clinicians through the importance weights assigned to each input variable by the attention layer.

Another approach in deep learning for sepsis prediction is to use temporally oriented Convolutional Neural Networks (CNNs). Lauritsen et al. [16] designed a CNN-RNN mixed architecture, which is a CNN followed by an RNN layer, to predict sepsis onset with data three hours prior to onset. Both Lauritsen et al. [17] and Moor et al. [22] proposed the use of Temporal CNN (TCN), a CNN architecture that uses causal convolutional so there is no data leakage from future observations to past ones similar to RNNs. It is well known that many healthcare diseases often progress over time. Therefore, RNNs and TCNs can provide better prediction performance by capturing the temporal progressing patterns. Therefore, we will adopt RNN in this work. However, we will also need to take each traumatic patient's profiles into consideration. Therefore, in this work, we apply a Multi-Modal RNN for early sepsis onset prediction.

One major problem in sepsis prediction, which has been less studied in the literature, is the class imbalance problem between sepsis and non-sepsis cases. As mentioned, many studies [12,13,17,19] create one sample per hospital admission. In this case, around 30% sepsis cases is not that rare. Typical machine learning techniques like over/under sampling [25] or class weighting [8] can help alleviate this level of imbalance. However, patients can be in different situations day by day. It is important to discriminate between days where sepsis is or is not present even for the same patient. In this case, the number of non-sepsis instances for learning can be dramatically increased and sepsis onset becomes a very rare event. In the literature, only one study by Ramchand [23] looked into addressing the rare event problem of predicting sepsis onset by using an ensemble of Long-

Short Term Memory (LSTM) RNNs. In this work, we propose to use pre-training as a novel method to address the serious class imbalance problem.

3 The Proposed Method

In this section, we describe our framework in three aspects, that is, early sepsis prediction setup, Multi-Modal RNN as the deep learning architecture, and the proposed pre-training technique to address our serious class imbalance problem.

3.1 Early Sepsis Prediction Setup

In this study, we utilize data collected from patients at night to predict whether the onset of sepsis will occur within the next 24 hours. Specifically data recorded between the hours of 10 p.m. and 6 a.m. to predict whether sepsis onset will occur in the next 24 hours from 7 a.m. to 6 a.m. the next day. This is a sliding window approach but for a fixed window of observation and a variable prediction horizon. While most studies opt for the fixed time before onset and sequential framing of onset prediction, we designed this methodological framing to be practical when deployed in clinical settings and best fulfill the needs of clinicians. We use features recorded during this nighttime window because 1) most staff is gone for the night so the model captures a gap in observation, 2) data collected at night is more likely to reflect the actual physiology of the patient as they are less exposed to external stimuli of hospital staff, 3) there are fewer interruptions to data collection as diagnostic procedures and interventions are typically planned during daytime hours, and 4) the night window is immediately adjacent to the period of time in which we predict sepsis onset. For these reasons, we believe that the nighttime window of data will be best suited for deployment and aid onsite staff in identifying the early stages of sepsis.

Given only the temporal features like body temperature and heart rate hour by hour, our sepsis prediction problem can be framed as a multivariate time-series classification task. However, some static features such as age and sex could be potential risk factors contributing sepsis. Therefore, each observation in our setting contains 9 hours of T temporal feature values and S static feature values as illustrated in Fig. 1. Based on this input, the multi-modal RNN model described in the next subsection predicts if the corresponding patient will develop sepsis in the next 24 hours as shown in Fig. 1.

3.2 Multi-Modal RNN

From our methodological setup, each instance has a 9xT matrix of temporal data and an S-dimensional vector of static data as input for sepsis prediction. To integrate these data, we propose a Multi-Modal RNN depicted in Fig. 2. The beginning of the architecture is separated into two components, temporal (blue) and static (red), that takes in the respective input.

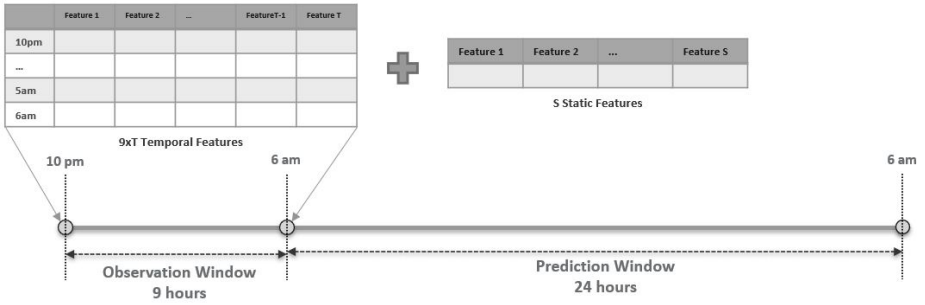

Fig. 1: Sepsis Prediction Setup. Temporal Features are extracted from the nighttime window between 10 p.m. and 6 a.m., forming a 9xT matrix for predicting sepsis onset within the subsequent 24 hour prediction window. Further adding data to inform the prediction, there is an additional S-dimensional vector of S static features which are consistent throughout the admission.

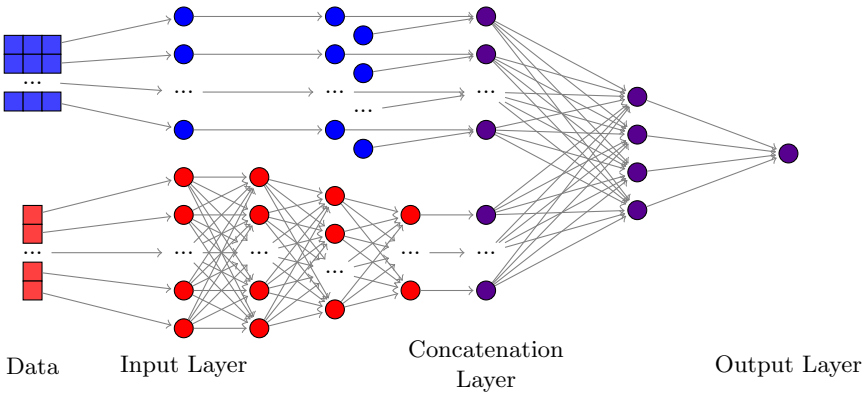

Fig. 2: Multi-Modal RNN Architecture. Temporal data is processed using a bidirectional GRU layer (blue) to capture temporal information while static data is processed using a deep dense layer architecture (red). Results from each are combined via concatenation then processed through further dense layers (purple).

The temporal component consists solely of a bidirectional RNN layer to capture the temporal information from the temporal features. Of the different recurrent units, Long-Short Term Memory (LSTM) [10] and Gated Recurrent Unit (GRU) [3] are often used because of their inclusion of a gating mechanism to solve the vanishing gradient problem that often afflicts other recurrent architectures like plain RNN. In studies comparing the two architecture in medical prediction tasks, GRU is preferable because of its simpler architecture and demonstrated better performance [4], which we found to be the case as well for sepsis prediction in our own testing. As well, some studies tested applying bidirectional RNN, an

RNN which uses the standard forward in time recurrent layer plus an additional backward in time feeding recurrent layer, to healthcare related prediction tasks to improve the performance beyond regular unidirectional RNN [20]. Therefore, we opt for a Bidirectional GRU layer for modeling temporal EHR data.

The 9xT matrix is fed into the network via the input layer which is of the same dimension. The input layer then forwards the data to the bidirectional GRU layer such that each hour of data is sent to one unit in the recurrent layer. Inside the recurrent layer, each unit expands the representation of the T features to a 256-dimensional vector such that the output dimension of this layer is 9x512; 9 recurrent units each producing a 512-dimensional vector. The hidden state of each recurrent unit is passed further down the network to incorporate information from each hour of the observation. As such, a flattening layer is used to flatten the hidden state from the recurrent layer to form a 4608-dimensional vector so it can be combined and processed with hidden state from the static component.

Next, static features are processed through a deep neural network comprised of an input layer, followed by three dense layers consisting of 16, 8, and 1 hidden unit(s) respectively. The results of each component are then combined via a concatenation layer which concatenates the hidden states together to form on 4609-dimensional vector that can then be passed through further dense layers to melds their respective information for prediction. Hence, the concatenation layer is followed by another dense layer consisting of 64 hidden units which passes onto the output with a single unit for binary classification. Moreover, each dense and recurrent layer is lead by Batch Normalization layers to aid in achieving model convergence. In this configuration, the Multi-Modal RNN accepts a 9xT matrix and an S-dimensional vector, then outputs a single scalar value between 0 and 1 which is interpreted as the probability for sepsis onset to occur. As such the probability is rounded such < 0.5 is 0, no onset, and ≥ 0.5 is 1, onset.

3.3 Pre-Training for Class Imbalance

Recent progress from pre-training shows that given an appropriate pre-trained model, a satisfied performance can be achieved with only a small amount of labeled examples from the target domain [14]. The observation inspires us to develop a down-sampling strategy for the class imbalance problem with a pre-trained model.

Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ denote a training data set where $y_i \in \{1, \ldots, C\}$ is the corresponding label for $\mathbf{x}_i$. A classification model can be learned by empirical risk minimization (ERM)

$$\min_\theta \frac{1}{n} \sum_i \ell(\mathbf{x}_i, y_i; \theta)$$

where θ is the parameter of the model and $\ell(\cdot)$ is a loss function where cross-entropy loss is popular in deep learning. For a C-class classification problem, the

task is equivalent to

$$\min_{\theta} \sum_{c}^{C} \frac{n_c}{n} f(\mathbf{x}, y, c) \tag{1}$$

where n_c shows the number of examples from the c-th class and

$$f(\mathbf{x}, y, c) = \frac{1}{n_c} \sum_{i:y_i=c} \ell(\mathbf{x}_i, y_i; \theta)$$

With the formulation, it is obvious that the loss will be dominated by the major class if there is some class j such that $\forall c, c \neq j, n_j \gg n_c$.

Then, we try to illustrate that a pre-trained model can help. By initializing the model with pre-trained weights, model will be fine-tuned with a small learning rate and the optimization problem becomes

$$\min_{\theta} \sum_{c}^{C} \frac{n_c}{n} f(\mathbf{x}, y, c) \quad s.t. \quad \|\theta - \theta_0\|_F \leq \gamma \tag{2}$$

The benefit of pre-trained model can be demonstrated in the following theorem.

Proposition 1 *A function f is L-Lipschitz continuous if*

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$$

Theorem 1 *Let $\theta(\mathbf{x})$ be L-Lipschitz in $\mathbf{x}$ and θ^* denote the solution of Eqn. 2, then if $\gamma \leq \frac{1}{8L}$ we have*

$$\forall i, j, \quad \|\theta^*(\mathbf{x}_i) - \theta^*(\mathbf{x}_j)\|_2^2 \geq \|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2 - \|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2/2 - 1/32$$

Proof.

$$\begin{aligned}
&\|\theta^*(\mathbf{x}_i) - \theta^*(\mathbf{x}_j)\|_2^2 = \|\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i) + \theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j) + \theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2 \\
&= \|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2 + \|\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)\|_2^2 + \|\theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j)\|_2^2 \\
&+2\langle\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i), \theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j)\rangle + 2\langle\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i), \theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\rangle \\
&+2\langle\theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j), \theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\rangle \\
&\geq \|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2 - 2\|\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)\|_2\|\theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j)\|_2 \\
&-2\|\theta^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)\|_2\|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2 - 2\|\theta_0(\mathbf{x}_j) - \theta^*(\mathbf{x}_j)\|_2\|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2
\end{aligned}$$

Due to the smoothness of $\theta(x)$, we have

$$\|\theta^*(\mathbf{x}_i) - \theta^*(\mathbf{x}_j)\|_2^2 \geq \|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2 - 2L^2\gamma^2 - 4L\gamma\|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2$$

Remark Theorem 1 illustrates that fine-tuning from a pre-trained model appropriately can preserve the diversity in the pre-trained representations, which can avoid the collapse of minor classes in the class imbalance problem. Note that many unsupervised representation learning adopts instance classification as the pretext task for pre-training [2, 9], we have

$$E_{i \neq j}[\theta_0(\mathbf{x}_i)^\top \theta_0(\mathbf{x}_j)] = 0$$

With the observation from instance classification, we have the bound as follows.

Corollary 1 *Let $\theta(\mathbf{x})$ be L-Lipschitz in $\mathbf{x}$ and θ^* denote the solution of Eqn. 2, then if $\gamma \leq \frac{1}{8L}$ and θ_0 is pre-trained with instance classification and $\|\theta_0(\mathbf{x})\|_2 = \|\theta^*(\mathbf{x})\|_2 = 1$, we have*

$$E_{i \neq j}[\theta^*(\mathbf{x}_i)^\top \theta^*(\mathbf{x}_j)] \leq 0.37$$

Proof. Due to Jensen's inequality, we have

$$E[\|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2] \leq \sqrt{E[\|\theta_0(\mathbf{x}_i) - \theta_0(\mathbf{x}_j)\|_2^2]} = \sqrt{2}$$

Then, we can observe the bound by taking it back to the inequality.

Explicitly, the examples from minor classes will not be dominated by those from the major class. Based on this theoretical analysis, we propose to initialize the weights of our multi-modal RNN by a pre-training task of instance classification, that is, each instance is uniquely identified.

More concretely, given a dataset of N instances, each instance is assigned a unique identifier such that there are N different classes which will be the target of the instance classifier. A multi-modal RNN is constructed with similar architecture to the binary multi-modal RNN classifier illustrated in Fig. 2 with the only change appearing in the output layer. As the prediction task is now multi-class classification, the output layer will consist of N units instead of 1. The instance classifier is then trained using all N instances until satisfactory accuracy ($> 95\%$) is achieved. Then, the weights except the output layer will be used as the initial weights of the binary Multi-Modal RNN before training.

4 Experiments

4.1 Data Description

We were provided de-identified EHR data by Harborview Medical Center (HMC), which pertains to patients aged 16 years and older admitted into the ICU following injury between the years of 2012 and 2019. In total, the data contains EHR data for 2,802 patients, 486 of which developed sepsis during their stay in the ICU, making up approximately 17%. Sepsis is defined according to the 2016 international guidelines, Sepsis-3 [28–30], as a clinically suspected infection

associated with worsening of organ dysfunction. We restricted to include only infections that were identified between hospital days 3 and 14, as this is a period of time when patients are at risk for hospital acquired infections [21] and still in the acute phase of critical illness. Patients were labeled as having sepsis if they meet the following two criteria within three days of each other.

- Suspected of having an infection. Confirmed with a positive culture sample.
- Exhibits worsening organ dysfunction denoted by a change in the Sequential Organ Failure Assessment (SOFA) [31] score greater than or equal to 2 ($\Delta SOFA \geq 2$).

The time of onset is determined by the time that the positive culture sample was drawn from the patient due to it being the most specific landmark for HMC clinicians' concern that an infection is present or developing.

The provided EHR data includes details about patients' demographics, information about their injuries, physiological signatures recorded hourly, therapeutics administered, and comorbidities. Comorbidity info was available in the data, but because it is not always available at the point of care, we did not include in our feature set. From this data, we constructed the features into subsets that we call layers. Each of these layers encapsulated one category of data identified as being a contributing factor to the development of sepsis. Layer 1 is physiology of the patient aggregated each hour, layer 2 contains patient factors and initial physiology from the first 48 hours, and finally layer 3 is cumulative exposures treated as events accumulated over time. Layers 1 and 3 are temporal, while layer 2 is static. The purpose of parsing the data into layers is that layers can be dynamically isolated and combined to test their contribution to prediction as well as their interaction with one another. Table 1 summarizes the detailed feature information for each layer.

EHR being a large, heterogeneous data format is susceptible to large amounts of missingness and sparsity. Another benefit of our prediction setup is a short observation window. In this case, it is often reliable to apply Last Observation Carry Forward (LOCF) to impute missing values, except for Mean Arterial Pressure which was calculated from Systolic and Diastolic Blood Pressure as shown in Eqn. 3 [27].

$$MAP = \frac{2(DBP) + SBP}{3} \tag{3}$$

4.2 Instance Extraction and Inclusion

After cleansing the data and parsing the features into layers, we extract instances from the nighttime window to train and evaluate our proposed method. For each visit, we extract the temporal features between the hours of 10 p.m. and 6 a.m. for each night the patient is in the ICU for days 3 through 14. Day 1 and 2 are excluded, since hospital acquired infections develop after 48-72 hours of hospitalization [21]. Infections before day 3 are considered to be acquired from the community or associated with healthcare; therefore it is not a problem of the

Layer	Features
Layer 1	Heart Rate
	Diastolic Blood Pressure (DBP)
	Mean Arterial Pressure (MAP)
	Respiratory Rate
	Temperature
	Fraction of Inspired Oxygen (FiO2)
Layer 2	Age
	Sex
	Mechanism of Injury
	Was Transferred from Another Hospital?
	Has Head Injury?
	First Systolic Blood Pressure in ED
	Reverse Shock Index
	Max Base Deficit*
	Max Lactate*
	Total Red Blood Cell Count*
	Total Crystalloids (Not in OR)*
	Apache II
	Antibiotic Exposure*
	Number of Surgeries*
	Emergency Department Disposition
Layer 3	IV Fluid Bolus Volume
	Red Blood Cell Count
	Vent Days
	Number of Surgeries
	Surgery Duration

Table 1: Feature layers. Layer 1 is physiological data aggregated hourly. Layer 2 is comprised of patient factors and initial physiology. Lastly, layer 3 is cumulative exposures which are events accumulated over time. * indicates that they are collected from and aggregated over the first 48 hours of the patient's admission.

trauma population. Windows after 14 days are also excluded as after 14 days, patients are no longer considered acutely ill but instead are considered to have chronic critical illness which has different phenotype from sepsis in the acute recovery phase. In addition, instances containing any null values after LOCF are also excluded. For each study/visit, the static features are identified and added to each instance.

Lastly, above instances are labeled according to if the first occurrence of sepsis onset occurs within 24 hours after the observation window. While sepsis recidivism is a big problem as well, our setup targets the first sepsis event for the purpose of early prediction. Therefore, instances after the first onset for septic patients were excluded. This results in a total of 25,952 instances with 471 positive ones and 25,481 negative ones, where the class imbalance problem can be observed.

4.3 Experimental Setup

To evaluate our proposed method, we compare the performance of multi-modal RNN with pre-training (named as 'Pre-trained Multi-Modal RNN') with that without pre-training (named as 'Multi-Modal RNN') and the state-of-art traditional machine learning method, XGBoost [1]. These methods are applied to four different feature sets (i.e., Layer 1 only, Layer 1 + Layer 2, Layer 1 + Layer 3, and Layer 1 + Layer 2 + Layer 3) for the purpose of contribution-factor analysis.

To evaluate each of these configurations, 5-fold stratified cross validation was used. In addition, due to the class imbalance between negative and positive instances, the majority class would dominate the empirical risk minimization during the training. In this case, the model would learn to always predict negative. Therefore, during the training of the sepsis vs. non-sepsis classification, we under-sampled the negative instances and over-sampled the positive instances for the training data only, each with 2,600 instances.

For Multi-Modal RNN experiments, numerical features were each scaled to [0, 1] using Min-Max scaling. Moreover, the static component of the Multi-Modal RNN is excluded from the architecture if no static features are included in the corresponding configuration. For XGBoost that expects the input to be a single vector, the temporal feature matrix was flatten to a vector and concatenated with the static features to form one long vector.

Lastly, each of these experiments were evaluated using area under the receiver operating characteristic curve (AUROC). AUROC is often preferred for evaluating binary classification results and is the most used metric to get a holistic understanding of the performance of the classifier. However, it does not provide any information about how the model is performing for positive and negative classes separately. Therefore, we also assess confusion matrix counts along with the sensitivity and specificity. These metrics are then aggregated across all five folds such that the ROC curve is the intercept of each curve and the confusion matrix is the sum of all the matrices.

4.4 Performance Comparison

Table 2 compares the performance of two baselines (i.e., Multi-Modal RNN and XGBoost) using different feature sets, where the highest two for each metric are underlined. First, it can be observed from the comparison between feature sets that accumulated features in Layer 3 can significantly improve the performance for both models. However, the contribution of static features is limited, which may indicate that temporal progressing pattern is more useful to predict sepsis. Second, Multi-Modal RNN with the ability to capture the temporal progressing pattern outperforms XGBoost that treats each feature independently, especially for positive cases. This confirms the effectiveness of temporal models on sepsis prediction as in the literature.

Considering that ICUs prefer to use machine learning models as a pre-screening assistant, they care more about missed positive cases (i.e., never being checked in-depth by clinicians), while a large amount of false positives can be

Feature Set	AUROC	True Positives	True Negatives	Sensitivity	Specificity
XGBoost					
Layer 1	0.706	287	17577	0.609	0.690
Layers 1+2	0.713	296	17161	0.628	0.673
Layers 1+3	0.740	315	17632	0.669	0.692
Layers 1+2+3	0.743	317	<u>17749</u>	0.673	<u>0.697</u>
Multi-Modal RNN					
Layer 1	0.712	319	16370	0.677	0.642
Layers 1+2	0.711	299	17698	0.635	0.695
Layers 1+3	<u>0.782</u>	<u>353</u>	17425	<u>0.724</u>	0.684
Layers 1+2+3	<u>0.780</u>	<u>330</u>	<u>18314</u>	<u>0.700</u>	<u>0.719</u>

Table 2: Performance of XGBoost and Multi-Modal RNN using each feature set. The highest two for each metric is underlined.

Model	AUROC	True Positives	True Negatives	Sensitivity	Specificity
XGBoost	0.740	315	<u>17632</u>	0.669	<u>0.692</u>
MMR	0.782	353	17425	0.724	0.684
Pre-trained MMR	<u>0.787</u>	<u>390</u>	15602	<u>0.828</u>	0.612

Table 3: Performance comparison using Layers 1 and 3 as the feature set. The best performance for each metric is underlined.

easily filtered according to the confidence score (i.e., probability to be positive). We compare the proposed Pre-trained Multi-Modal RNN (shortened to Pre-trained MMR) to the configuration with the highest true positives (i.e., Layer 1 + Layer 3 using MMR). Table 3 shows the comparison. We observe that using a using a sufficiently trained instance classification model to pre-train our proposed Multi-Modal RNN provides the highest AUROC and improves the sensitivity of the model in predicting the target event significantly. The sensitivity increases 10% when using pre-training which is a meaningful improvement for the purpose of pre-screening, although sacrificing the specificity that is less critical. This demonstrates the effectiveness of pre-training for very rare events or a very minor class.

5 Conclusion

In this study, we propose a new prediction setup for early sepsis onset anticipation by machine learning, which is more applicable and deploy-able as a pre-screening assistant in ICUs. To use both temporal and static features as potential contributing factors to sepsis, we develop a Multi-Modal RNN model pre-trained with a sufficiently accurate instance classifier that is proposed to solve the serious class imbalance problem. The proposed methodology demonstrates promising results in enhancing model sensitivity in early prediction of sepsis which will assist ICU staffs in intervening early. Using data about patient

physiology and cumulative summation of exposures, our Multi-Modal RNN exhibited favorable performance over XGBoost. As well, pre-training the model using the weights from instance classification further increased the sensitivity of the proposed model in capturing sepsis onset.

However, the high sensitivity comes from sacrificing of the specificity, although the highest AUROC is achieved. One potential factor is the over and under sampling procedure adopted, where certain negative ones have been ignored. Therefore, a future work is to eliminate the over and under sampling, but still addressing the class imbalance problem using a new loss function inspired by our theoretically analysis from pre-training. Besides, one potential reason that Layer 2 static features did not help too much is our deep neural network design with only 1-dimensional output to be contacted with the temporal features, which can be explored in the future. Another critical future direction is to enable the interpretability of the model.

References

1. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020), http://proceedings.mlr.press/v119/chen20j.html
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association **24**(2), 361–370 (2017)
5. Eguia, E., Cobb, A.N., Baker, M.S., Joyce, C., Gilbert, E., Gonzalez, R., Afshar, M., Churpek, M.M.: Risk factors for infection and evaluation of sepsis-3 in patients with trauma. The American Journal of Surgery **218**(5), 851–857 (2019)
6. Eriksson, J., Eriksson, M., Brattström, O., Hellgren, E., Friman, O., Gidlöf, A., Larsson, E., Oldner, A.: Comparison of the sepsis-2 and sepsis-3 definitions in severely injured trauma patients. Journal of Critical Care **54**, 125–129 (2019)
7. Futoma, J., Hariharan, S., Heller, K.: Learning to detect sepsis with a multitask gaussian process rnn classifier. In: International Conference on Machine Learning. pp. 1174–1182. PMLR (2017)
8. He, J., Cheng, M.X.: Weighting methods for rare event identification from imbalanced datasets. Frontiers in big Data p. 108 (2021)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 9726–9735. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00975, https://doi.org/10.1109/CVPR42600.2020.00975

10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. Kaji, D.A., Zech, J.R., Kim, J.S., Cho, S.K., Dangayach, N.S., Costa, A.B., Oermann, E.K.: An attention based deep learning model of clinical events in the intensive care unit. PloS one **14**(2), e0211057 (2019)
12. Khojandi, A., Tansakul, V., Li, X., Koszalinski, R.S., Paiva, W.: Prediction of sepsis and in-hospital mortality using electronic health records. Methods Inf Med **57**(04), 185–193 (2018)
13. Khoshnevisan, F.: A Variational Recurrent Adversarial Multi-Source Domain Adaptation Framework for Septic Shock Early Prediction Across Medical Systems. North Carolina State University (2021)
14. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V. Lecture Notes in Computer Science, vol. 12350, pp. 491–507. Springer (2020). https://doi.org/10.1007/978-3-030-58558-7_29, https://doi.org/10.1007/978-3-030-58558-7_29
15. Kumar, A., Roberts, D., Wood, K.E., Light, B., Parrillo, J.E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al.: Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine **34**(6), 1589–1596 (2006)
16. Lauritsen, S.M., Kalør, M.E., Kongsgaard, E.L., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B.: Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artificial Intelligence in Medicine **104**, 101820 (2020)
17. Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B.: Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications **11**(1), 1–11 (2020)
18. Lauritsen, S.M., Thiesson, B., Jørgensen, M.J., Riis, A.H., Espelund, U.S., Weile, J.B., Lange, J.: The consequences of the framing of machine learning risk prediction models: evaluation of sepsis in general wards. arXiv preprint arXiv:2101.10790 (2021)
19. Liu, R., Greenstein, J.L., Granite, S.J., Fackler, J.C., Bembea, M.M., Sarma, S.V., Winslow, R.L.: Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the icu. Scientific reports **9**(1), 1–9 (2019)
20. Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., Gao, J.: Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1903–1911 (2017)
21. Monegro, A.F., Muppidi, V., Regunath, H.: Hospital acquired infections. In: StatPearls [Internet]. StatPearls Publishing (2020)
22. Moor, M., Horn, M., Rieck, B., Roqueiro, D., Borgwardt, K.: Temporal convolutional networks and dynamic time warping can drastically improve the early prediction of sepsis. arXiv preprint arXiv:1902.01659 (2019)
23. Ramchand, S.: An ensemble lstm for rare event detection (2020)
24. Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Ñamendys Silva, S.A., Martin-Loeches, I., Leone, M., Lupu, M.N., Vincent, J.L.: Sepsis in intensive care unit patients: Worldwide data from the intensive care over nations audit. Open forum infectious diseases **5**(12), ofy313–ofy313 (2018)

25. Sasada, T., Liu, Z., Baba, T., Hatano, K., Kimura, Y.: A resampling method for imbalanced datasets considering noise and overlap. Procedia Computer Science **176**, 420–429 (2020)
26. Scherpf, M., Gräßer, F., Malberg, H., Zaunseder, S.: Predicting sepsis with a recurrent neural network using the mimic iii database. Computers in biology and medicine **113**, 103395 (2019)
27. Sesso, H.D., Stampfer, M.J., Rosner, B., Hennekens, C.H., Gaziano, J.M., Manson, J.E., Glynn, R.J.: Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. Hypertension **36**(5), 801–807 (2000)
28. Seymour, C.W., Liu, V.X., Iwashyna, T.J., Brunkhorst, F.M., Rea, T.D., Scherag, A., Rubenfeld, G., Kahn, J.M., Shankar-Hari, M., Singer, M., et al.: Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). Jama **315**(8), 762–774 (2016)
29. Shankar-Hari, M., Phillips, G.S., Levy, M.L., Seymour, C.W., Liu, V.X., Deutschman, C.S., Angus, D.C., Rubenfeld, G.D., Singer, M.: Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA : the journal of the American Medical Association **315**(8), 775–787 (2016)
30. Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., et al.: The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama **315**(8), 801–810 (2016)
31. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P.M., Thijs, L.G.: The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. Intensive care medicine **22**(7), 707–710 (1996)
32. Wong, A., Otles, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., et al.: External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine **181**(8), 1065–1070 (2021)