Sets of Sub-Sequences based Sepsis Prediction for ICUTrauma Patients

Sijin Huang

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Ankur Teredesai

Juhua Hu

Katherine Stern

Program Authorized to Offer Degree:

Computer Science and Systems

University of Washington

Abstract

Sets of Sub-Sequences based Sepsis Prediction for ICU Trauma Patients

Sijin Huang

Chair of the Supervisory Committee:
Ankur Teredesai
School of Engineering and Technology

Sepsis is an extreme inflammatory response of the body to an infection. It is one of the leading causes of death in ICUs worldwide, resulting in approximately 25% mortality in critically ill populations. Early identification and intervention are crucial to reducing sepsis-associated mortality and improving patient prognosis because severe sepsis cases can lead to organ failure and other life-threatening complications. Diagnosis of Sepsis is challenging in terms of diagnostic accuracy and timeliness due to ambiguous symptoms and individual differences. In recent years, numerous efforts have been made using machine learning methods for sepsis prediction. However, there are still very limited successful implementations due to limitations in consistency of data input, which cannot fit the characteristics of uncertain time intervals and large number of missing values in real-world scenarios. In this study, we propose an innovative

approach to predict sepsis occurrence in real-time, using a flexible graph structure to model patient health records, predicting future sepsis risk at any time after the first 48 hours using observations data from the past 12 hours. To the best of our knowledge, our proposed approach is the first ever implementation that uses a graph representation to overcome the problem of irregular input features and continuous risk prediction thereby improving the compatibility of the model in clinical settings. Experiments on multi-year longitudinal data from a large level-1 trauma center demonstrate the effectiveness of our approach.

1 INTRODUCTION

Sepsis is a high mortality condition in ICU which leads to organ failure, and requires immediate healthcare interventions [15, 21, 24, 28]. According to a global audit conducted in 2018 [21], the incidence of sepsis in ICU hospitalized patients varied from 13.6% to 39.3% of all hospital patients in different regions. ICU mortality of sepsis patients was 25.8% and in-hospital mortality was 35.3% in sepsis patients, both significantly higher than non-septic patients.

Early intervention can improve prognosis of patients [10]. Taking antibiotics can prevent sepsis from developing, or reduce symptoms when an infection begins. For each hour of antibiotic intervention, mortality was reduced by 15% [9]. Therefore, early and accurate diagnosis of sepsis is crucial to the prognosis of patients. While there exists general guidance for sepsis treatment [20], sepsis is challenging to detect early in the critically ill trauma population because organ dysfunction from injury can mask or obscure clinical signs of infection [2, 4]. Physicians still lack personalized systems for decision support.

Given the importance of early identification of sepsis, many prior researcher efforts have tried to use machine learning models to predict the occurrence of sepsis ahead of its onset in patients based on data from Electronic Health Records (EHR). Various studies use traditional machine learning models [7, 14], such as random forest, XGboost, with EHR data in a fixed time window as input. The drawback of these methods is the lack of EHR data modeled as a time series and the inability to extract temporal patterns. Some newer methods use deep learning models [6, 14, 23] such as LSTM and CNN to directly learn temporal features. Liu et. al. [14] demonstrated the superiority of the LSTM method over traditional models through experimental comparisons. However, these new methods have very high data quality requirements such as: (a) the input time series must be at a fixed sampling interval, and (b) records with missing values are not allowed as inputs when using these models for scoring a patient. However, in clinical practice, it is difficult to ensure such high data quality. On the contrary, some vital sign data are collected by sensors, and the update frequency may only be a few minutes, while some laboratory test data or interventions such as surgery only occur once within a few days. Furthermore, patients may leave the ICU due to diagnostic imaging, surgical intervention, etc. At this time, data cannot be continuously collected, and missing values are unavoidable. Therefore, the existing models ignore the difference of sampling intervals in the real scenarios, and at the same time have to discard some features that do not meet the input requirements of the models.

In our study, we creatively use a flexible graph data representation and Graph Neural Networks (GNN) to overcome these challenges. The graph structure can express time intervals between feature records, and the model has the characteristics of naturally supporting unstructured data input, a capability that greatly increases the implementation flexibility.

In the experimental design process, we have also further improved the setup. Many of the earlier studies using machine learning models [1, 7, 11, 14, 26, 27] were retrospective and identified the timestamp of the sepsis onset event, then record input data by looking back for a fixed time interval. Only one input window

is selected for each patient, the experimental conclusions of these studies are not sufficient to demonstrate clinical utility when frequent and regular predictions are needed for a single patient. Motivated by the unmet need for early detection tools which can predict sepsis prospectively, and therefore translate into the clinical setting that in each hour, based on data from the past 12 hours, the risk of sepsis happening in next few hours was predicted so as to realize the variable prediction window.

Through more reasonable experiments, compared with baseline LSTM model, our model can flexibly process EHR data with missing values and has higher prediction performance. Due to pragmatic data utilization of data, our GNN model is promising for the early detection of sepsis.

The main contributions of this thesis can be summarized as:

- Design and develop the first ever medical Graph Neural Network (GNN) based sepsis prediction system
- Demonstrate that Graph representation helps support features with regular time intervals and mitigate the loss due to missing values.
- First continuous rolling window prediction setup respective of time of day or hour. i.e. use previous 12 hours to predict risk of developing sepsis in next few hours (variable output)

2 RELATED WORK

In this section, I will first review some related work about the related studies in sepsis prediction. Then, I will introduce general graph neural networks (GNN) for learning flexible relations and its successful usage.

2.1 Machine Learning for Early Sepsis Prediction

Traditional machine learning models such as XGboost and Random Forest [1, 7, 11, 26] have been applied to sepsis prediction. Barton el al. [1] tested XGBoost for up to 48h sepsis prediction with 6 vital signs as input. Khojandi et al. [7] used Random Forest to predict sepsis and mortality risk based on first 12, 24 and 48 hours after patients' admission to ICU. Tang et al. [26] applied PCA and SVM to classify the severity of sepsis. The challenge of these studies is that traditional machine learning models rely on hand-designed features and cannot model the continuous transition of input data over time. These weaknesses limit the expressive ability and predictive performance of traditional models.

Deep neural networks have also recently been applied to sepsis prediction due to their automatic feature extraction and capability to model time series. Through experimental comparison, Liu et al. [14] proved that the RNN model has better results in the early prediction of sepsis than Generalized Linear Model and XGBoost. Moor et al. [17] extended the Temporal CNN model to analyze EHR data using causal convolution. These deep models have good performance because they can capture slowly changing patterns over time. However, they require temporal features be input in a fixed sampling interval format, and missing values are not allowed. In clinical practice, the recording frequencies of different features are not the same, and missing values are inevitable. These challenges limit the clinical usefulness of these models.

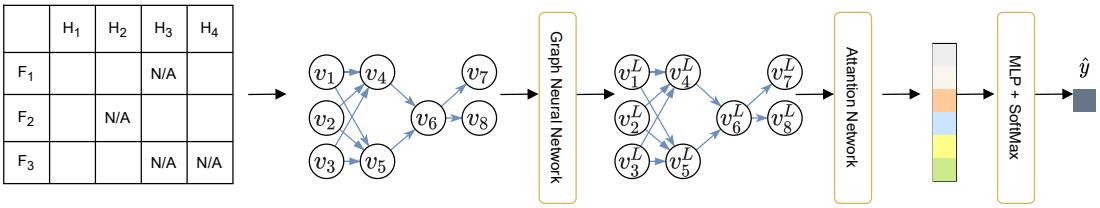


Fig. 1. Workflow of our sepsis prediction model

2.2 Common Evaluation Experimental Designs

In the relevant literature, the most common classification prediction target of machine learning models is whether sepsis will occur in a fixed time in the future. Khojandi et al [7] used input data 12, 24, and 48 hours before onset for prediction. Van et al. [27] used a time window collected 6 hours before onset to predict the probability of developing sepsis.

When collecting experimental data, these studies are based on the known onset time and retrieve previous data in a fixed-length duration. For patients with sepsis, only the latest time window was collected into the dataset. This experimental setup can only be used for retrospective studies with known onset times. Such a setting does not evaluate the model's false positive predictions in the early stages of patients, and it is difficult to be confidently applied directly to real medical infrastructures. In our experimental setting, sepsis onset probabilities were predicted every hour, and the precision of both positive and negative predictive outcomes was assessed, more in line with the metrics that clinicians care about.

2.3 Graph Neural Networks

Graph structures can flexibly describe unstructured relationships between objects in the real world. To extract information from graphs, graph neural networks (GNNs) are proposed. Recently, GNNs have been applied in various fields, including traffic prediction [5], antibacterial discovery [25], recommendation systems [29], etc.

In recommendation systems, there is a group of GNN models [18, 30–32] that predict next action (e.g. which item to click) of a user by modeling user historical behaviors. This type of model uses a graph structure to express user behavior history, and uses GNN models to extract transition patterns of historical events. These studies demonstrate the feasibility of analyzing temporal events using GNNs. However, these models only support input sequences of events with an explicit order, and ignore the length of time intervals between events. In this study, We extended these methods to better support the characteristics of EHR data.

3 PROPOSED METHOD

In this section, we will introduce the pipeline of our proposed method. Figure 1 shows the main workflow of our prediction model.

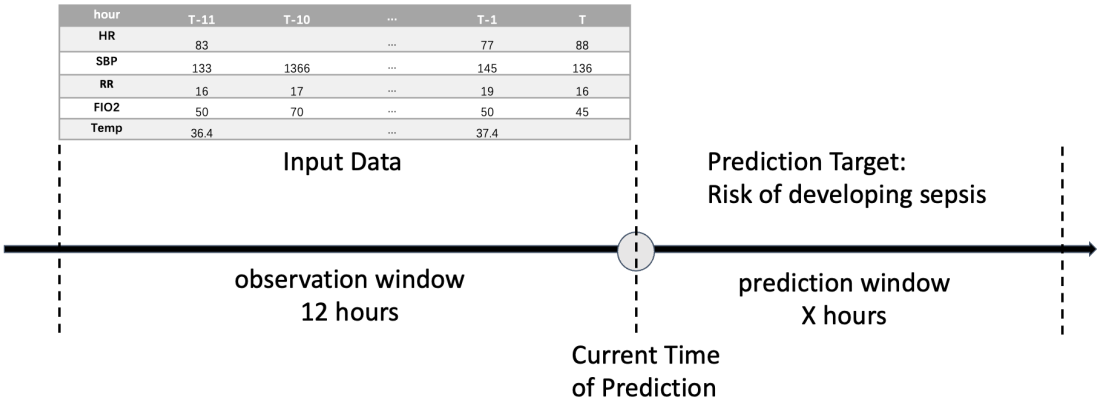


Fig. 2. Input Data and Prediction Setup

3.1 Prediction Setup

We tend to train a model to make real-time predictions for sepsis onset risk. The overall setup for input data and prediction target is shown in Figure 2. At any time, the input data is the recorded temporal features in previous 12 hours, and the model could output the risk of sepsis onset in next certain period of hours. The 12-hour input data is called the observation window, and the certain period of hours in future is called the prediction window. With the same observation window, our model can predict sepsis risk in various prediction windows.

We are treating our task as a binary classification problem. In order to conduct and evaluate early prediction of a patient's first sepsis event, we excluded data after the initial onset. For training data generation, a sliding window with a length of 12 hours is used to extract the input data, and the prediction target would be whether the patient developed sepsis during the prediction window.

Prediction window is a hyper-parameter which will influence the amount of positive windows. Since the amount of negative input windows are far more than positive windows, random sampling is used to decrease the impact of negative windows to our model.

3.2 Feature Groups

We divided all temporal features into different groups.

Vital Signs are the physiological measurements of patients, which serve as the baseline of our selected features. Vital Signs include heart rate (beats per minute), blood pressure (systolic, diastolic, mean arterial pressure measured in mmHg), respiratory rate (breaths per minute), temperature (degrees celcius) and fraction of inspired oxygen (%FiO2).

hourTally	21	22	23	24	25	26	27	28	29	30	31	32
HR	78	84	79	80	77	85	80	78	88	78	87	89
Rolling avg HR						80.500	80.833	79.833	81.333	81.000	82.667	83.333
Delta HR							-0.500	-2.833	8.167	-3.333	6.000	6.333
Trend HR									**Increase**		**Increase**	**Increase**

Fig. 3. Example of calculating trends: $\Delta HR_{27} = HR_{27} - Avg(HR_{21}, \cdots, HR_{26})$

Cumulative Exposures are interventions accumulated over time, including intravenous (IV) fluid bolus volume (Liters in excess of 0.5 given within 1 hour), units of red blood cells transfused, days exposed to invasive mechanical ventilation (i.e., ventilator days), surgeries (count), and surgery duration (hours).

Laboratory Data are laboratory test results including serum bicarbonate (mEq/L), strong ion difference (mEq/L), blood urea nitrogen (mg/dL), creatinine (mg/dL) and white blood cell count (x10^9 cells/L).

3.3 Data Preprocessing

3.3.1 Analyzing Trends of Vital Signs. Changes in physiologic parameters (such as an increase in respiratory rate, heart rate, decline in blood pressure, and an increase or decrease in temperature) can be signs of an evolving infectious process, but the strength and direction of these physiologic signals may differ between patients depending on age and baseline comorbidities. Thus, the trends of vital signs may foreshadow the future sepsis development. We analyzed trends as supplemented information for all features in vital sign group. At each hour t, the trend is defined as the difference between current feature value f_t and the rolling mean of previous 6 hours $Avg(f_{t-6}, f_{t-5}, ..., f_{t-1})$. Figure 3 shows an example of calculating trends for heart rate.

3.3.2 Restore Cumulative Exposures. Features in Cumulative Exposures group are recorded as accumulative values over time. This recording format is beneficial for machine learning methods since continuous exposures may cause higher risk of sepsis, and can also cover the great amount of missing data when intervention is not implemented. However, we aim to model the transition of temporal recordings, in this step we restore the accumulated values to incremental values as shown in Figure 4, to express emergence of interventions at certain time, which is consistent with record format of Vital Signs or Laboratory Data. When the accumulated value has no change from previous record, we determine such exposure or intervention is not happening in this hour, and would be marked as blank and treated as missing in the following steps.

3.3.3 Features Categorization. Graphs are capable of representing relations between discrete nodes. When using graph structure to model EHR data, it is necessary to convert continuous features into categories. Clinician experts helped us to formulate the categorizing criteria. The categorization thresholds for Vital Signs and Trends are described in Table 1 and standards for Laboratory Data are shown in Table 2. For each

hour	46	47	48	49	50	51	52
cumulative IV fluid bolus	0	0	0.5	0.5	0.5	1	1
IV bolus delta			0.5			0.5	
cumulative surgHours	0	0	0	2	2	2	2
Surg receiving				1	1		
vent	0	0	0	1	1	1	1
Vent receiving				1	1	1	1

Fig. 4. Example of Cumulative Exposures restoration

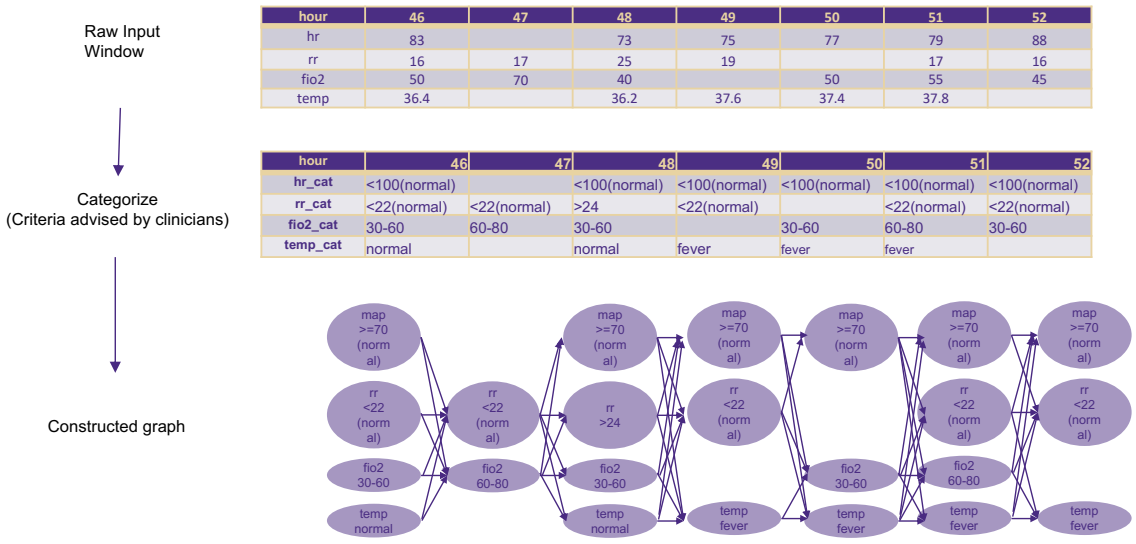


Fig. 5. Proposed method to construct graph from EHR

feature of Cumulative Exposures, there is only one emergence category. We removed the hourly dosage differences because the recordings are too sparse, and integration of labels could help the model to learn a more generalized representation for these exposures or interventions.

Figure 5 displays the method of feature categorization. Every hour in the input 12-hour window, each feature with numeric record is categorized into discrete levels. Features with missing data is kept blank in this step. In this way, each feature is a category.

3.3.4 Constructing Temporal Graphs. After the categorization of continuous temporal features, the input window becomes a table of discrete values where each row represents one hour and the columns are different features. Then, it can be modeled as a directed graph. The overall design can be found in Figure 5. Node v_i in the graph denotes the category of one healthcare feature. To express hourly transitions, all node

Table 1. Categorize thresholds and significant trends for Vital Sign features

Feature	Categories	Significant Trends associated with infection/sepsis
Heart Rate	>120 111-120 100-110 <100 consider normal	Increase by 10
Systolic Blood Pressure	>100 consider normal 90-100 low <90 very low	Decrease by 10
Diastolic Blood Pressure	>60 <60 low	Decrease by 10
Mean Arterial Blood Pressure	>70 consider normal 65-70 <65	Decrease by 10
Respiratory Rate	>24 22-24 12-21 consider normal <12	Increase by 5
Fraction Inspired Oxygen (%)	>80 60-80 30-60 <30 consider normal	Increase by 10
Temperature	>38 high 36-38 consider normal <36 low	Up one category

Table 2. Categorize thresholds for Layer4 features

Feature	Bicarbonate	Strong ion difference	Blood urea nitrogen / creatinine ratio	White blood cell count
Categories	1 = >22 2 = 20-22 3 = 17-19 (bad) 4 = <17 (very bad)	1 = >22 2 = 20-22 3 = 17-19 (bad) 4 = <17 (very bad)	1 = <5 (low) 2 = 5-20 (normal) 3 = 20-40 (elevated) 4 = >40 (high)	1 = >14 (high) 2 = 12-14 (elevated) 3 = 4-11 (normal) 4 = <4 (low - also not good)

pairs from two successive hours are connected, thus each edge (v_i, v_j) explicitly represents the transition from the state v_i in previous hour to the state v_j in current hour.

3.4 Graph Representation Extraction

3.4.1 Learning Event Embedding. Each healthcare event node v_i is embedded into a latent vector $\mathbf{v_i} \in \mathbb{R}^d$, where d is the embedding dimension. Latent vector $\mathbf{v}$ would be learnt from the GNN. The GNN model could capture graph features from complex edge relations, which is suitable for learning dynamic transitions.

Node vectors would be updated through Gated GNN as the following equations:

$$a_{s,i}^t = A_{s,i:} \left[v_1^{t-1}, .., v_n^{t-1}\right]^T H + b \tag{1}$$

$$z_{s,i}^t = \sigma\left(W_z a_{s,i}^t + U_z v_i^{t-1}\right) \tag{2}$$

$$r_{s,i}^t = \sigma\left(W_r a_{s,i}^t + U_r v_i^{t-1}\right) \tag{3}$$

$$\widetilde{v_i^t} = tanh\left(W_o a_{s,i}^t + U_o\left(r_{s,i}^t \odot v_i^{t-1}\right)\right) \tag{4}$$

$$v_i^t = \left(1 - z_{s,i}^t\right) \odot v_i^{t-1} + z_{s,i}^t \odot \widetilde{v_i^t} \tag{5}$$

In equation 1, A is the adjacency matrix of observation graph. $a_{s,i}$ is the aggregation of neighbor representations. $z_{s,i}$ and $r_{s,i}$ are reset gates and update gates in standard GRU equations respectively.

Vectors of neighbors propagate to the current node. The final output state of current GNN layer is combination of the previous hidden state and the candidate state, controlled by update gate and reset gate. The gated update equation can be also written as:

$$v_i^t = GRU\left(v_i^{t-1}, a_{s,i}^t\right) \tag{6}$$

3.4.2 Generating Observation Window Embedding. One comprehensive representation is generated for the whole observation window. This global representation is dynamically composed of all node embeddings in the graph through an Attention Network. Specifically, the observation window embedding considers complete 12-hour feature sequences and also pays extra attention to the latest records.

The input window could be jointly represented with global vector s, which is generated considering all events in the observation window by an Attention Network.

Node vectors are gathered after feeding forward process of GNN. The local state embedding s_l is simply defined as the average of embeddings happen at last hour $s_l = \overline{v_{i,n}}$.

Since states in a observation period could have different impact on sepsis onset, we take the soft-attention mechanism to capture global observation embedding s_g.

$$\alpha_i = q^T \sigma\left(W_1 s_l + W_2 v_i + c\right) \tag{7}$$

$$s_g = \sum_{i=1}^{n} \alpha_i v_i$$

where s_g is the weighted sum of all state embeddings.

Finally, a linear transformation is used on both local and global observation embedding to generate the hybrid observation embedding s_h:

$$s_h = W_3 \left[s_1; s_g \right] \tag{8}$$

3.5 Making prediction and model training

After generating observation embedding s_h, we can evaluate the sepsis onset risk z_i by simply dot product each candidate outcome embedding a_i with observation embedding s_h:

$$\hat{y}_i = s_h^T a_i \tag{9}$$

For training our GNN model, the loss function on one session is defined as the binary cross entropy between predicted score and ground truth sepsis onset label.

4 EXPERIMENTS AND ANALYSIS

4.1 Data Description

The data for experiments are collected from Harborview Medical Center in Seattle, WA between January 2012 and December 2019. The cohort consists of 2802 severely injured adults over 16 years old, admitted to the trauma ICU who required at least 72 hours of invasive mechanical ventilation. Patients whose first hospital unit of admission was a location other than the ICU or an emergent procedural unit (e.g., operating room, angiography) were excluded. In the cohort, 486(17%) of patients developed sepsis during the stay in ICU. The EHR data contains demographics, injury details, physiological records and therapeutic interventions. Since hospital acquired infections develop after 48-72 hours of hospitalization [16] and sepsis after injury peaks during the first 1-2 weeks of admission, we focus on sepsis events that developed between hospital days 3 through 14. Table 3 shows some demographic characteristics of our dataset.

Post-traumatic sepsis was classified retrospectively using a combination of clinical criteria and chart review. We used the CDC's adult sepsis surveillance criteria [19] with a priori modifications utilizing readily obtainable EHR data to improve specificity for the trauma population. We required that all of the following be present: 1) an order for a new IV or qualifying oral antibiotic, not administered within the previous 48 hours and excluding antibiotics used for surgical prophylaxis, 2) a body tissue culture was ordered within 48 hours of antibiotic initiation, 3) a qualifying antibiotic was sustained for at least 4 consecutive days, or until death or discharge, and 4) a 2-point increase in the maximum daily sequential organ failure assessment (SOFA) score occurred within 3 days before and 3 days after the qualifying culture. Sepsis was confirmed

Table 3. Characteristic of patients

		All patients (N = 2802)			
		Non-sepsis (N = 2316)		Sepsis (N = 486)	
Characteristics		Count	(%)	Count	(%)
Gender	Male	1698	73.3	380	78.2
	Female	618	26.7	106	21.8
Age	16-29	542	23.4	106	21.8
	30-39	311	13.4	82	16.9
	40-49	301	13	64	13.2
	50-59	402	17.4	89	18.3
	60-69	354	15.3	75	15.4
	70+	406	17.5	70	14.4

Table 4. Random sampling for training data

Prediction Window	Window Label	Training Set		Testing Set
		without sampling	after sampling	
12h	Pos	4704 (0.8%)	23520 (18.4%)	1098 (0.8%)
	Neg	524867	104194	132001
24h	Pos	9425 (1.7%)	47125 (31.3%)	2195 (1.6%)
	Neg	520062	103233	130872
48h	Pos	18510 (3.4%)	92550 (47.7%)	4285 (3.2%)
	Neg	511061	101433	128814

by chart review in the follow subgroups: culture negative sepsis and patients meeting partial but not full clinical criteria.

4.2 Sliding Window Extraction

We use a 12-hour sliding window to generate input data from EHR data. More than 660,000 total windows are extracted. Table 4 shows that there exists of serious data imbalance challenge. To relieve the imbalanced windows, random negative down-sampling and positive up-sampling is applied to the training set. We randomly delete 20% of all negative windows, and duplicate all positive windows 4 times. After sampling, the positive ratio for 24-hour prediction increased from 1.7% to 31.3%, alleviating the imbalance. Besides, weighted cross entropy is also used to pay more attention on positive samples while training.

4.3 Evaluation Metrics

Most early studies [1, 3, 8] of deploying machine learning models on sepsis onset prediction use Area Under the Receiver Operating Characteristic curve (AUROC) as major evaluation metric. Operating characteristic curve (ROC) is plotted with true positive rate and false positive rate at different confidence threshold settings. AUROC is suitable for binary classification tasks, which can measure the rank correlation between

predicted probabilities and targets. Also because the considers both positive and negative samples equally, it can also serve as a mixed measurement of sensitivity and specificity.

However, our major concern is that AUROC may be biased when evaluating severe imbalanced data. In our collected data, positive samples are far less than negative samples. True positive rate (i.e., sensitivity) could be precisely measured by AUROC, but false positive rate would be imprecise due to large amount of negative samples. Area Under the Precision-Recall Curve (AUPRC) shows the trade-off between precision and recall rate across different decision thresholds. Recall rate is exactly same as true positive rate, meanwhile precision score is more decisive on imbalanced data. In this work, we choose AUPRC as our major metric.

Besides AUPRC and AUROC, other common metrics for binary classification including F-1 score, precision rate, recall rate, sensitivity and specificity are also recorded to measure diverse model performance.

4.4 Parameters Setup

Following similar GNN methods [12, 13], we set the dimension of latent vectors $d = 100$ for all nodes. Besides, we tune other hyper-parameters using 5-fold cross validation. Orthogonal initialization [22] is used to initialize parameters for GRU cells in the Readout function, based on its good performance on RNN cells. All other parameters are initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. We choose the Adam optimizer for training, where the learning rate is set to 10^{-4} and will decay by 0.5 every 20 epochs. Moreover, the batch size is set to 100 and the L2 penalty with 10^{-4} is added to avoid over-fitting.

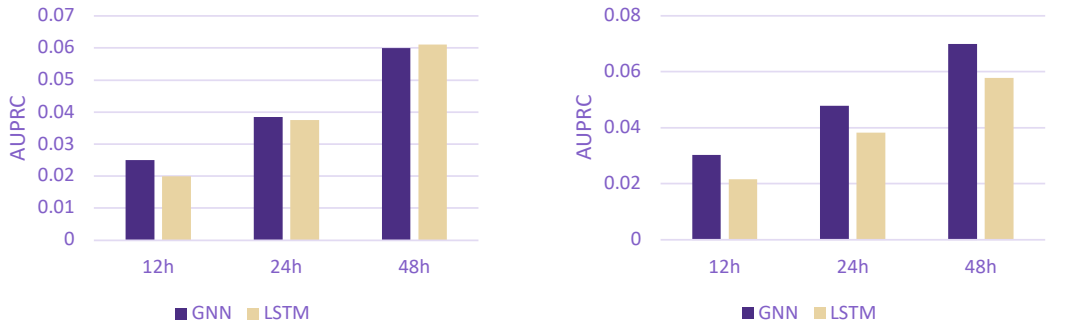
4.5 Comparison with Baseline Methods

To demonstrate the performance of the proposed approach, we first compare with other commonly used deep learning based sepsis prediction models. LSTM is chosen as the baseline method, for its wide usages in sepsis prediction and other tasks like time series analysis.

LSTM can accept input both continuous variables and discrete variables. Since feature categorization using expert-designed criteria is a step in our data processing pipeline, we want to compare our complete approach with LSTM on continuous features, to prove both the validity of categorization and strength of GNN-based model.

For a fair comparison, LSTM and GNN have exact same input windows and prediction targets. Identical features are selected, and Last Observation Carried Forward (LOCF) imputation is applied to both methods, in order to relieve the restriction of LSTM that missing values are not allowed. Due to the limitation that some features in Cumulative Exposures and Laboratory Data have too many missing values and cannot implement imputation, these features are excluded when comparing with baseline model. Two feature selection setups are used: a) Vital Signs with LOCF imputation, b) Vital Signs with LOCF imputation + Trends.

The AUPRC score between different methods are shown in Figure 6. When using feature groups of Vital Signs with imputation, our proposed GNN model performs as good as baseline LSTM in terms of



(a) Features: Vital Signs + LOCF Imputation

(b) Features: Vital Signs + LOCF Imputation + Trends

Fig. 6. AUPRC comparison between LSTM and GNN

Table 5. 24h prediction Metrics comparing LSTM

Model	AUROC	specificity	sensitivity	precision	AUPRC	F1
GNN	0.71	0.92	0.29	0.06	0.05	0.09
LSTM	0.64	0.93	0.21	0.05	0.04	0.08

AUPRC. After including Trends features, our GNN model has a significant improvement in AUPRC than LSTM model. This indicates that our method has utilization ability at least comparable to LSTM, and will even outperforms LSTM in some input feature groups. To further inspect the prediction results, Table 5 reports more detailed metrics for the 24-hour prediction window setup in Figure 6b. GNN has a slight improvement in terms of precision score, but has an obvious increase in terms of sensitivity, which domains the improvement in overall AUPRC score.

4.6 Benefit of severely missing features

The proposed GNN-based method is flexible in constructing temporal transitions of various features, which naturally supports modeling features with missing records. As shown in Table 6, there are some features in Cumulative Exposures and Laboratory Data are extremely rare, which only recorded in less than 5% of all the hours during the admission in ICU. Data imputation cannot reliably process these features, thus, traditional time-series models are unable to leverage information from these features. In this section, we compare the prediction performance between baseline Vital Signs features and adding Cumulative Exposures and Laboratory Data with missing values, in order to test the ability of our method to capture patterns of features even with extreme missing values, and measure the incremental value that these rare feature types have for model performance.

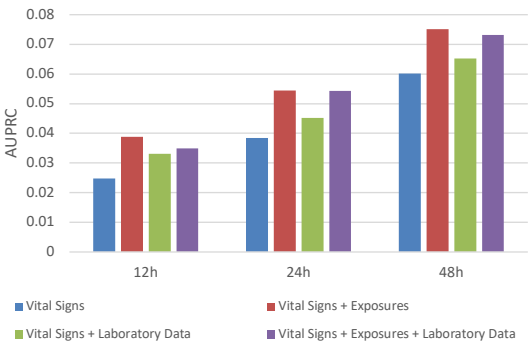


Fig. 7. AUPRC comparison for adding features with missing values

Figure 7 reports AUPRC when adding Cumulative Exposures and Laboratory Data among three prediction window setups. A clear improvement in terms of AUPRC can be observed when adding Cumulative Exposures or Laboratory Data. Cumulative Exposures brings the most significant increase, which matches the expectation since these exposures are highly related to sepsis development.

This experiment verified the capability of our approach to accept features with severe missing data, and proved the benefit of taking Cumulative Exposures and Laboratory Data into account.

4.7 Comparing Combinations of Feature Groups

In order to analyze the impact of combining different groups of features in our method, further experiments are applied to study the prediction performance according to differentiate feature selection settings.

The candidate feature groups are Vital Signs with or without LOCF data imputation, Trends of Vital Signs, Cumulative Exposures and Laboratory Data. We proposed different combinations of feature groups, and evaluated the performance of each combination in 12, 24 and 48-hour prediction window setups, in order to analyze the supportive factors of each feature group. Figure 8 demonstrates the AUPRC metric on various of feature groups combinations.

First we compare the influence of data imputation for vital signs. Features in vital signs have approximately 30% of missing values, and can be relieved by data imputation techniques, which is also widely used in other studies for sepsis prediction. Though our GNN-based method is compatible with data containing missing values, here we tried to examine the impact of imputation solely to Vital Signs as well as other feature groups.

From Figure 8 we can observe that for models that only use vital signs as input, adding data imputation would cause a decrease in terms of AUPRC in all prediction windows. This fact may indicate that missed features could also serve as contributing feature related to sepsis onset. The missing data for Vital Signs mainly come from when the patient left ICU due to interventions including surgeries or diagnostic imaging, which may be related to risks of developing sepsis. Table 6 shows that there is less missingness among

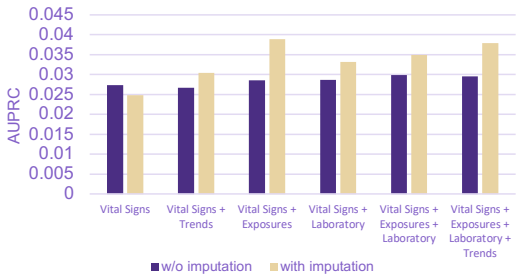
Table 6. The proportion of missing values for each feature

		All patients (N = 2802)	
		Non-sepsis (N = 2316)	Sepsis (N = 486)
Feature Group	Feature	Missing Values (%)	Missing Values (%)
Vital Signs	Heart rate	31%	10%
	Systolic BP	37%	21%
	Mean BP	36%	20%
	Diastolic BP	37%	21%
	Respiratory rate	33%	14%
	Temperature	59%	40%
	FiO2	66%	39%
Exposures	IV Fluid bolus volume (L)	98%	98%
	Cumulative sum RBCs	99%	99%
	Surgeries	98%	98%
	Vent days	50%	19%
Laboratory Data	Blood urea nitrogen	30%	17%
	Creatinine	31%	17%
	White blood cell count	31%	17%
	Lymphocyte count	97%	95%
	Neutrophil count	97%	95%
	Urine output (mL)	72%	69%

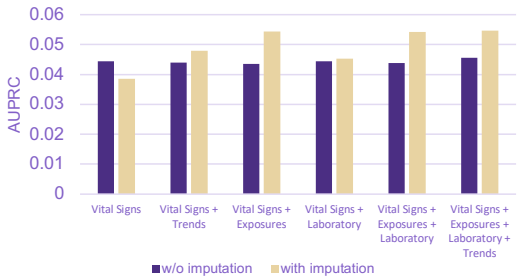
patients who developed sepsis, possibly because they are not clinically well and being more closely monitored. Missing data is not completely at random in this setting and rather than artificially masking this fact with an imputation strategy, the GNN is able to use missingness as a piece of information in its own right. When missingness is not masked over, performance improves, suggesting it is valuable for prediction.

Another fact according to Figure 8 is that without data imputation for Vital Signs, adding more feature groups such as Cumulative Exposures or Laboratory Data only slightly affect the prediction performance. However, when including imputed values for Vital Signs, adding more feature groups would have a significant improvement, demonstrating that the imputation of vital sings has positive contribution to utilize other feature groups. It's possible that the greater number of Vital Sign nodes resulting from imputation allows for more relationships between Vital Signs and other feature types to be identified and integrated into the model.

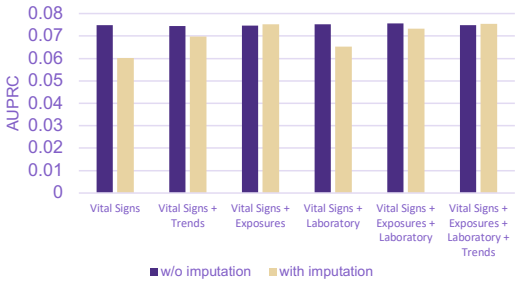
We then compared model performance when different feature groups were included: Vital Signs with or without LOCF data imputation, Vital Sign Trends, Interventions and Laboratory Values. Figure 9 demonstrates the AUPRC score with different prediction windows and selected features. The features of Vital Signs with imputation + Trends + Cumulative Exposures have the best prediction performance among all prediction windows. When limited to only two feature groups, Vital Signs with imputation + Cumulative Exposures gives the best prediction result. A detailed evaluation with more metrics for our best prediction model is reported in Table 7.



(a) 12-hour prediction window



(b) 24-hour prediction window



(c) 48-hour prediction window

Fig. 8. AUPRC comparison with Vital Signs imputation

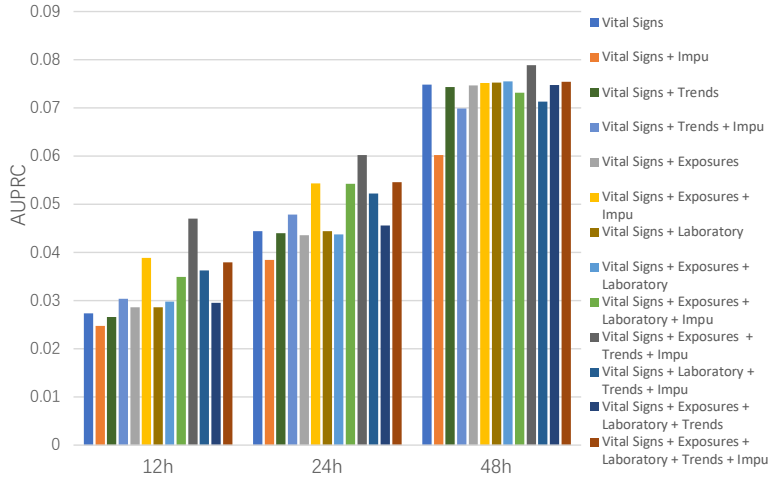


Fig. 9. AUROC comparison among component combinations

Table 7. Best prediction metrics

Predict window	AUROC	specificity	sensitivity	precision	AUPRC	F1
12h	0.8146	0.9687	0.2437	0.0611	0.0470	0.0976
24h	0.7786	0.8777	0.4332	0.0561	0.0603	0.0993
48h	0.7339	0.7009	0.6333	0.0657	0.0789	0.1191

4.8 Comparing GNN Layers

In this section we analyze the influence of the number of GNN layers. The number of GNN layers control the distance of the node-level message propagation process in the model. At the Lth layer in the model, the nodes in the graph can receive information from at most L-hop neighbors. According to the directed graph construction method, each edge represents the feature transition between two successive hours, thus L-hop neighbor only include healthcare records within L+1 hours. Limit a GNN model with L layers will restrict the model to focus on transition patterns within certain time period of L+1 hours, and a deeper model with more GNN layers would be able to catch longer term patterns.

In the experiment, we tested the AUPRC prediction performance for GNN with 1, 2, 4, or 6 layers under different prediction window. The best feature groups combination is used (i.e. Vital Signs with imputation + Trends + Cumulative Exposures).

Experiment results can be found in Figure 10. In general, a deeper model would decrease the prediction performance in terms of AUPRC in all prediction windows. Besides, the decrease in more notable when predicting a shorter 12-hour length window size, and the influence of GNN layers can be hardly observed for a 48-hour prediction window. Overall, model with 2 GNN layers have the best performance across different prediction windows, indicating that patterns within 3 hours may be sufficient to generate node representation in our method.

These findings may be explained by the model relying more on patterns extracted from acute changes in physiology or clinical events. Studies [1, 8] show the feasibility to predict sepsis onset based on input vital signs data within a 3-hours period, therefore the short-term patterns may be the dominant factor of our GNN model. The rapid decrease in terms of AUPRC of a shorter prediction window may indicate that short-term forecast relies more on swift changes within 3 hours, but a longer changing pattern may be equivalently supportive for 48-hour prediction.

5 CONCLUSION

Sepsis is a very urgent condition in ICU with a high mortality. Machine learning models can help predict the possibility of ICU patients developing sepsis according to the historical data. However, in real medical scenarios, lots of treatments and experiments happen without a certain pattern but leading to a very sparse data set. It is crucial for machine learning models to handle the missing values properly.

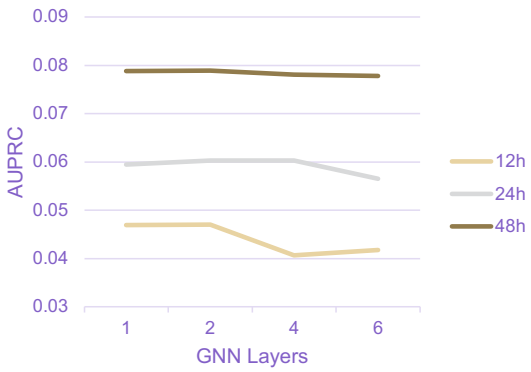


Fig. 10. AUPRC comparison among GNN layers

We proposed a GNN model supporting sparse input data including high proportion of missing values for sepsis prediction. Experimented on real-world ICU data, our approach outperforms baseline LSTM model with a 13% increment in terms of AUPRC. Furthermore, by adding different features to the model like Cumulative Exposures and Laboratory Data, GNN model shows the capability to guarantee the prediction performance over different prediction windows. This suggests that GNN, with its pragmatic use of clinical data, may have promising applications for the early detection of sepsis in critically ill populations. GNN model can be extended to improve the personalized healthcare interventions and provide reasonable treatment suggestions.

REFERENCES

[1] Christopher Barton, Uli Chettipally, Yifan Zhou, Zirui Jiang, Anna Lynn-Palevsky, Sidney Le, Jacob Calvert, and Ritankar Das. 2019. Evaluation of a Machine Learning Algorithm for up to 48-Hour Advance Prediction of Sepsis Using Six Vital Signs. *Computers in Biology and Medicine* 109 (June 2019), 79–84. https://doi.org/10.1016/j.compbiomed.2019.04.027

[2] E Cole, S Gillespie, P Vulliamy, K Brohi, and Organ Dysfunction in Trauma (ORDIT) study collaborators. 2020. Multiple organ dysfunction after trauma. *Br. J. Surg.* 107, 4 (March 2020), 402–412.

[3] Hong-Fei Deng, Ming-Wei Sun, Yu Wang, Jun Zeng, Ting Yuan, Ting Li, Di-Huan Li, Wei Chen, Ping Zhou, Qi Wang, et al. 2021. Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. *Iscience* (2021), 103651.

[4] Emanuel Eguia, Adrienne N Cobb, Marshall S Baker, Cara Joyce, Emily Gilbert, Richard Gonzalez, Majid Afshar, and Matthew M Churpek. 2019. Risk factors for infection and evaluation of Sepsis-3 in patients with trauma. *Am. J. Surg.* 218, 5 (Nov. 2019), 851–857.

[5] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 922–929.

[6] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. 2019. An attention based deep learning model of clinical events in the intensive care unit. *PLoS One* 14, 2 (Feb. 2019), e0211057.

[7] Anahita Khojandi, Varisara Tansakul, Xueping Li, Rebecca S Koszalinski, and William Paiva. 2018. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods Inf. Med.* 57, 4 (Sept. 2018), 185–193.

[8] Norawit Kijpaisalratana, Daecha Sanglertsinlapachai, Siwapol Techaratsami, Khrongwong Musikatavorn, and Jutamas Saoraya. 2022. Machine Learning Algorithms for Early Sepsis Detection in the Emergency Department: A Retrospective Study. *International Journal of Medical Informatics* 160 (April 2022), 104689. https://doi.org/10.1016/j.ijmedinf.2022.104689

[9] Richard Y Kim, Alex M Ng, Annuradha K Persaud, Stephen P Furmanek, Yash N Kothari, John D Price, Timothy L Wiemken, Mohamed A Saad, Juan J Guardiola, and Rodrigo S Cavallazzi. 2018. Antibiotic timing and outcomes in sepsis. *The American Journal of the Medical Sciences* 355, 6 (2018), 524–529.

[10] Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, David Gurka, Aseem Kumar, and Mary Cheang. 2006. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* 34, 6 (June 2006), 1589–1596.

[11] Simon Meyer Lauritsen, Bo Thiesson, Marianne Johansson Jørgensen, Anders Hammerich Riis, Ulrick Skipper Espelund, Jesper Bo Weile, and Jeppe Lange. 2021. The Consequences of the Framing of Machine Learning Risk Prediction Models: Evaluation of Sepsis in General Wards. *arXiv preprint arXiv:2101.10790* (2021).

[12] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[13] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: Short-Term Attention/Memory Priority Model for Session-based Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom). 1831–1839.

[14] Ran Liu, Joseph L. Greenstein, Stephen J. Granite, James C. Fackler, Melania M. Bembea, Sridevi V. Sarma, and Raimond L. Winslow. 2019. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Scientific Reports* 9, 1 (1 Dec. 2019). https://doi.org/10.1038/s41598-019-42637-5 Funding Information: We would like to thank Drs. Kathrine Henry, Suchi Saria, Nauder Faraday, and Adam Sapirstein for valuable discussion. Work supported by NSF EECS 1609038 and NIH UL1 TR001079. Publisher Copyright: © 2019, The Author(s).

[15] Nicasio Mancini. 2015. *Diagnostic Methods and Protocols*. Springer.

[16] Alberto F Monegro, Vijayadershan Muppidi, and Hariharan Regunath. 2020. Hospital acquired infections. In *StatPearls [Internet]*. StatPearls Publishing.

[17] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. 2019. Temporal convolutional networks and dynamic time warping can drastically improve the early prediction of sepsis. *arXiv preprint arXiv:1902.01659* (2019).

[18] Ruihong Qiu, Jingjing Li, Zi Huang, and Hongzhi YIn. 2019. Rethinking the Item Order in Session-Based Recommendation with Graph Neural Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 579–588. https://doi.org/10.1145/3357384.3358010

[19] Chanu Rhee, Zilu Zhang, Sameer S Kadri, David J Murphy, Greg S Martin, Elizabeth Overton, Christopher W Seymour, Derek C Angus, Raymund Dantes, Lauren Epstein, David Fram, Richard Schaaf, Rui Wang, Michael Klompas, and CDC Prevention Epicenters Program. 2019. Sepsis surveillance using adult Sepsis Events simplified eSOFA criteria versus sepsis-3 Sequential Organ Failure Assessment criteria. *Crit. Care Med.* 47, 3 (March 2019), 307–314.

[20] Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, et al. 2017. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine* 43, 3 (2017), 304–377.

[21] Yasser Sakr, Ulrich Jaschinski, Xavier Wittebole, Tamas Szakmany, Jeffrey Lipman, Silvio A Ñamendys-Silva, Ignacio Martin-Loeches, Marc Leone, Mary-Nicoleta Lupu, Jean-Louis Vincent, et al. 2018. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. In *Open forum infectious diseases*, Vol. 5. Oxford University Press US, ofy313.

[22] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120* (2013).

[23] Matthieu Scherpf, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder. 2019. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput. Biol. Med.* 113, 103395 (Oct. 2019), 103395.

[24] Christopher W Seymour, Foster Gesten, Hallie C Prescott, Marcus E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M Terry, and Mitchell M Levy. 2017. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine* 376, 23 (2017), 2235–2244.

[25] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 4 (2020), 688–702.

[26] Collin HH Tang, Paul M Middleton, Andrey V Savkin, Gregory SH Chan, Sarah Bishop, and Nigel H Lovell. 2010. Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. *Physiological measurement* 31, 6 (2010), 775.

[27] Franco Van Wyk, Anahita Khojandi, and Rishikesan Kamaleswaran. 2019. Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. *IEEE journal of biomedical and health informatics* 23, 3 (2019), 978–986.

[28] Maja von Cube, Martin Schumacher, and Jéan-Francois Timsit. 2020. Sepsis. *The Lancet* 396, 10265 (2020), 1804.

[29] Menghan Wang, Yujie Lin, Guli Lin, Keping Yang, and Xiao-ming Wu. 2020. M2GRL: A multi-task multi-view graph representation learning framework for web-scale recommender systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2349–2358.

[30] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 346–353.

[31] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. 2019. Graph Contextualized Self-Attention Network for Session-based Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 3940–3946. https://doi.org/10.24963/ijcai.2019/547

[32] Feng Yu, Yanqiao Zhu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2020. TAGNN: Target Attentive Graph Neural Networks for Session-based Recommendation. *CoRR* abs/2005.02844 (2020). arXiv:2005.02844 https://arxiv.org/abs/2005.02844