Recent Temporal Pattern Mining for Septic Shock Early Prediction

Farzaneh Khoshnevisan*, Julie Ivy §, Muge Capan ¶,Ryan Arnold ‖, Jeanne M. Huddleston **, and Min Chi‡
*‡ Department of Computer Science, North Carolina State University, Raleigh, NC, USA
§ Department of Industrial Engineering, North Carolina State University, Raleigh, NC, USA
¶ Drexel University's LeBow College of Business, Philadelphia, PA, USA
‖ Drexel College of Medicine, Philadelphia, PA, USA
** Mayo Clinic, Rochester, MN, USA
Email: *fkhoshn@ncsu.edu, §jsivy@ncsu.edu, ¶Muge.Capan@drexel.edu, ‖ryanarnold08@mac.com,
** jeannehuddleston@mac.com, and ‡mchi@ncsu.edu

***Abstract*—Sepsis is a leading cause of in-hospital death over the world and septic shock, the most severe complication of sepsis, reaches a mortality rate as high as 50%. Early diagnosis and treatment can prevent most morbidity and mortality. In this work, Recent Temporal Patterns (RTPs) are used in conjunction with SVM classifier to build a robust yet interpretable model for early diagnosis of septic shock. This model is applied to two different prediction tasks: visit-level early diagnosis and event-level early prediction. For each setting, this model is compared against several strong baselines including atemporal method called Last-Value, six classic machine learning algorithms, and lastly, a state-of-the-art deep learning model: Long Short-Term Memory (LSTM). Our results suggest that RTP-based model can outperform all aforementioned baseline models for both diagnosis tasks. More importantly, the extracted interpretative RTPs can shed some lights for the clinicians to discover progression behavior and latent patterns among septic shock patients.**

I. Introduction

Sepsis is the leading cause of mortality in the United States and the most expensive condition associated with in-hospital stay, accounting for 6.2% (nearly $24 billion) of total hospital costs in 2013 [1]. In particular, *Septic shock*, the most advanced complication of sepsis due to severe abnormalities of circulation and/or cellular metabolism [2], reaches a mortality rate as high as 50% [3] and the annual incidence keeps rising [4]. It is estimated that as many as 80% of sepsis deaths could be prevented with early diagnosis and intervention; indeed prior studies have demonstrated that *early diagnosis* and treatment of septic shock can significantly decrease patients' mortality and shorten their length of stay [5]–[7]. Formerly, multiple complex patient health scoring systems have been defined and employed for early diagnosis and early intervention of sepsis, such as SOFA score [8], MEDS [9], APACHE II [10], and PIRO score [11]. Despite these efforts, awareness of sepsis, and specially septic shock remains low.

One major challenge associated with early diagnosis of sepsis/septic shock is its subtle progression. The clinical signs and symptoms at the early stages of sepsis/septic shock are often subtle and unspecific. For example, only minor changes are reflected on the white blood cell count and body temperature at the early stages of sepsis. Moreover, infection, a hallmark of sepsis, is highly likely to progress to other disease and hence not a symptom exclusive to sepsis. Therefore, it is critical to learn about the discriminative interpretable patterns of sepsis/septic shock and capture informative progression during a patient's stay. In this work, our goal is to integrate Electronic Health Records (EHRs), clinical expertise, and machine learning to provide a *robust yet interpretable* data-driven framework for accurate early diagnosis of septic shock.

EHRs are multivariate time series data that typically contain noisy, sparse, and irregularly timed observations. For example, during a patient's visit, the body temperature is often measured a few times a day, while the white blood cells are only measured every other day. Because of this sparsity and irregular sampling, common methods for classification and prediction of multivariate time series data, such as similarity measures, and time series feature extraction methods, like discrete Fourier transform and discrete wavelet transform [12] cannot be directly applied to EHRs. In recent years, deep learning models specially Recurrent Neural Networks (RNNs) and RNN-based models such as LSTM [13] and gated recurrent unit (GRU) [14], have been shown to achieve the state-of-the-art results in many real-word applications with multivariate time series data including EHRs. Such models enjoy several nice properties such as strong prediction performance through deep hierarchical feature construction as well as the ability to effectively capture long-term temporal dependencies in time series data. Despite their extensive applications and great success, a large number of variables in EHRs sampled in irregular time point and time intervals can pose enormous challenges for deep learning. More importantly, these deep learning models are often treated as "black box" models because of the lack of interpretability – they are particularly difficult to understand because of their non-linear nature.

On the other hand, Temporal Sequential Pattern-based approaches are designed to extract interpretable yet meaningful temporal patterns directly from irregularly sampled multivariate time series data. In recent years, significant efforts have been made to develop and apply various pattern-based

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00033

approaches to EHRs [15], [16]. Generally speaking, these approaches can be divided into two stages: the Temporal Abstraction (TA) stage and the frequent pattern mining stage. TA transforms a series of raw-data time points into an interval-based temporal relations among meaningful clinical concepts, from which frequent pattern mining finds significant, interval-based temporal patterns. While such approaches have shown to be effective in inferring meaningful temporal patterns directly from EHRs, their performance has not been investigated on progressive disease such as septic shock and have not been directly compared against the state-of-the-art approaches such as deep learning.

In this work, we will directly compare one Temporal Sequential Pattern-based approach, Recent Temporal Patterns (RTPs) [15] against a series of baseline methods including atemporal baseline called Last-Value used in the original paper [15], six classic machine learning classifiers including: Logistic Regression and SVM which are widely used in analyzing EHRs, and a state-of-the-art deep learning method: Long Short-Term Memory (LSTM). More specifically, they will be compared on two different early diagnosis tasks for septic shock: *the visit level early diagnosis* and *the event level early prediction*. Overall, we show that RTPs can outperform all baseline models including LSTM on both early diagnosis tasks and more importantly, RTPs can indeed discover interpretable yet meaningful temporal patterns that would be informative for clinicians.

Our main contributions are summarized as follows.

(1) To the best of our knowledge, this is the first attempt to apply a Temporal Sequential Pattern-based method, Recent Temporal Patterns (RTPs) to a progressive disease – septic shock and compared it against several baseline models including deep learning.

(2) We run extensive experiments on evaluating RTP and various baseline models on both *the visit level early diagnosis* and *the event level early prediction* tasks while most prior research mainly focused on one or another but not both.

(3) We identify interpretable yet meaningful temporal patterns which can be informative for the clinicians and thus can be used as septic shock indicators.

The remaining parts of this paper are organized as follows. In Section II, we review related works. Section III presents the pattern mining approach used throughout this paper. In Section IV, we discuss experimental setup and introduce early prediction models and baselines. Section V presents the results. Finally, Section VI concludes the paper.

II. Background

A. Sequential Pattern-based Classification for EHRs

Shahar proposed knowledge-based temporal abstraction in 1997 [17], which transforms the time-stamped data into interval-based presentation. Such temporal abstraction provided a foundation for all the subsequent sequential pattern mining approaches applied to multivariate temporal data. Several groups explored combination of knowledge-based TA with Apriori-like algorithms to extract informative sequential patterns [18]–[20]. These algorithms can have high computational cost, especially when applying to large EHR datasets. Therefore, a number of studies have focused on improving the efficiency of pattern mining algorithms [15], [21], [22]. However, while these patterns can be informative, some of the extracted patterns may not be relevant to the classification task [23]. In the interest of avoiding huge redundant patterns, Batal et al., proposed Minimal Predictive Temporal Patterns framework to extract informative, non-spurious patterns and demonstrated this approach improves the performance of event prediction task [24].

The use of frequent temporal patterns as features for classification of multivariate time series, especially in EHR data, has been proven effective by numerous studies [24]–[26]. For example, KarmaLego algorithm was proposed by Moskovitch et al. as a fast pattern mining algorithm [22], which in a later study they adapted them as features for learning the classifier and clinical procedures prediction. They demonstrated that this prediction model has a significant improvement over static features [27].

Application of sequential pattern-based classification for chronic and long-developing disease prediction has been investigated by several studies, especially in diabetes domain [28], [29]. In a recent study, Orphanou et al. employed temporal association rules combined with naive Bayes classifiers for coronary heart disease diagnosis. They showed that using higher levels of temporal abstraction improves the prediction performance when long sequences and distant events are critical [16]. However, few studies implemented Temporal Sequential Pattern-based approaches for prediction of fast-progressive disease such as sepsis/septic shock. Recently, Ghosh et al. adapted temporal patterns resulting from contrast pattern mining approach in conjunction with Coupled Hidden Markov Model (CHMM) to predict septic shock prior to onset. Their achieved accuracy performance is 0.85 for 30 minutes and 60 minutes early prediction of septic shock. They demonstrated that this performance is significantly higher than applying SVM directly on data, or using CHMM with continuous multivariate data [30].

Event detection among sepsis/septic shock patients is highly sensitive to time and recency of the events, since changes in condition of patients can suddenly happen and result in more severe stages or eventual death. Therefore, we need to apply a mining algorithm that takes the most recent measurements into consideration, because they are more predictive of future events, rather than distant measurements. Recent Temporal Pattern Mining algorithm presented by Batal et al., is an efficient method that takes maximum gap constraint into account to control the recency of the patterns and combines them with SVM classifier [15]. We adapted this algorithm in this study to extract the patterns and build classifiers for early prediction of septic shock.

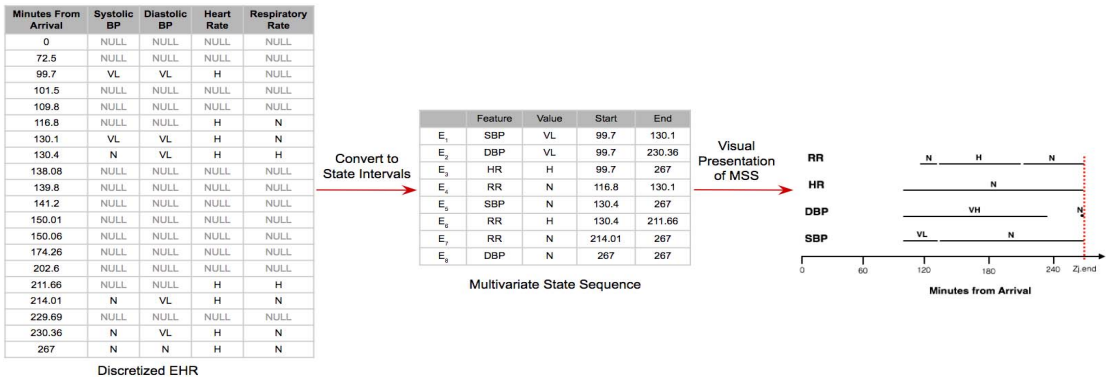


Fig. 1: An example of conversion from discretized data into Multivariate State Sequence with four temporal features.

B. Deep Learning for EHRs

In recent years, extensive research have shown the strength of deep learning models, especially Recurrent Neural Network (RNN) or RNN-based models such as LSTM and Gated Recurrent Unit (GRU) for modeling EHRs. Che et al. showed that GRU can capture long-term dependencies in data as well as taking informative missingness of the values into account in that it achieved better prediction performance than classic machine learning algorithms such as logistic regression and SVM [31]. Doctor AI uses GRU for multi-label prediction of future disease and shows the generalizability of this model among different medical centers with moderately high performance [32]. In another study they applied GRU for early prediction of heart failure and showed GRU outperforms all the classic machine learning models, such as SVM, KNN, and MLP [33]. Finally, Lipton et al. applied LSTM to EHRs and showed that LSTM can significantly outperform several strong baselines including MLP [34]. Generally speaking, GRU is a variation of LSTM [13]: GRU has a less complex structure, fewer parameters and is computationally more efficient than LSTM. Previously, Bahdanau et al. showed that GRU's performance is on a par with LSTM [35]. In this study, we explored both and found that LSTM achieves better performance. Thus, in the following we will focus on LSTM.

C. Septic Shock Prediction

Various classic machine learning algorithms have been applied to several public and private EHRs to predict sepsis-related outcomes. Such algorithms include Decision Tree [36], Recursive Partitioning And Regression Tree (RPART) [37], Support Vector Machines [38], [39], Neural Networks [40], Bayesian Networks [41], and a number of multivariate Logistic Regressions [42]–[44]. Additionally, [45] applied a LSTM based framework using two levels of imperfect yet informative labels to jointly predict septic shock. However, as far as we know, none of prior research has investigated the robustness of their models on both *the visit level early diagnosis* and *the event level early prediction*.

III. Recent Temporal Pattern Mining

Our dataset can be represented as $X = \{x_1, x_2, ..., x_N\}$ where N is the total number of hospital visits. It is composed of multivariate irregular time series data in that a visit x_i consists of a sequence of events: $x_i = \{x_i^1, ..., x_i^{T_i}\}$, where x_i^t represents the patient's records at time step t in x_i. We have $x_i^t \in \mathbb{R}^D$, where D is the number of attributes/features recorded at each event and T_i is number of events in the visit i which varies with different visits. Each x_i is associated with a visit-level label, denoted as $Y = \{y_1, y_2, ..., y_N\}$, where $y_i \in \{0, 1\}$. The objective in this study is to learn a function $f : X \rightarrow Y$ that can label unlabeled instances. We employ the following four steps to build such classifier:

1) Convert numeric time-series variables into time interval sequences using Knowledge-based TA.
2) Extract frequent recent temporal patterns from different classes of data (i.e. 0 and 1).
3) Transform each x_i into a fixed-size binary feature vector v_i, where the size of vector corresponds to the number of frequent RTPs from Step 2.
4) Build the classifier on the binary matrix generated in Step 3 to predict outcome.

A. Step 1. Knowledge-based Temporal Abstraction

This step can be divided into the following four sub-steps:
1. Data Discretization To discritize numeric values, we use 10%, 25%, 75%, and 90% percentiles of all values for each temporal feature in X, and define 5 categories of very low (VL), low (L), normal (N), high (H), and very high (VH). The left table in Fig. 1 shows a real-world sample of resulted discretized records including four temporal features.
2. Transforming into Multivariate State Sequence: The middle table in Fig. 1 demonstrates how the discretized data can be converted into a **Multivariate State Sequence (MSS)** z_i where each row presents a state interval (E) extracted from the visit's records. More specifically, we denote **State** S as (F, V), where F is a temporal feature and V is the discritized value for feature F at a given time and **State Interval** E is denoted as (F, V, s, e), where s and e refer to *start* and *end* time of the state (F,V). For example, the first row in

the middle table $E_1 = (SBP, VL, 99.7, 130.1)$ states that the patient's Systolic BP is very low from minutes 99.7 to 130.1.

Thus, we can convert each visit's data x_i into a corresponding MSS z_i by sorting all the state intervals by their start times:

$$z_i = \langle E_1, E_2, ..., E_n \rangle : E_j.s \leq E_{j+1}.s, j \in \{1, ..., n-1\}$$

Note that we also define $z_i.end$ as the end of the last state interval in MSS, i.e. $E_n.e$. For example, the right figure of Fig. 1 is a visualization of the MSS z_i and $z_i.end$ is 267. Applying this method on all $x_i \in X$ will transform X into a set of MSSs: $Z = \{z_1, z_2, ..., z_N\}$.

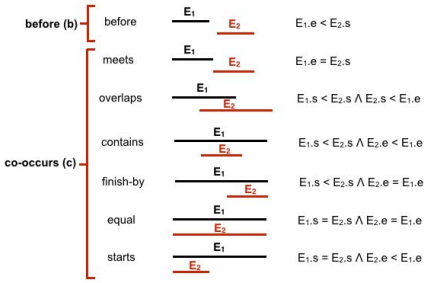


Fig. 2: Allen's seven basic temporal relations. Additionally, on the left, the two general relations in this study are presented.

3. Defining Temporal Relations: Allen [46] defined 7 basic temporal relations between states depending on their start and end time (Fig. 2). Here, we grouped the last six relations into one relation *co-occurs*, since they are slightly different from each other. Thus, two temporal relations between two instantaneous events are defined as follows:

- E_i **before** (b) E_j: When E_i ends before the start of E_j ($E_i.e < E_j.s$).
- E_i **co-occurs** (c) with E_j: When E_i and E_j have some overlap time ($E_i.s \leq E_j.s \leq E_i.e$).

4. Defining Temporal Patterns: Temporal patterns are generated by combining states and two aforementioned types of temporal relations: b or c to describe temporal dependencies in data. More specifically, for n states $\langle S_1, ..., S_n \rangle$, we define its corresponding temporal pattern: $P = (\langle S_1, ..., S_n \rangle, R)$, where R is an upper triangular matrix of relations: each cell $R_{i,j}$ in R corresponds to the relation between S_i and S_j, while $R_{i,j} \in \{\text{b,c}\}$. Additionally, the size of the temporal pattern P is determined by the number of states S it contains. As an example, suppose that we want to present a 5-pattern with three temporal features: Temperature (T), Systolic Blood Pressure (SBP), and Lactate (L) as shown in Fig. 3. Each state is an abstraction of a variable and the half matrix on the right side shows the temporal relations between each pair of states. For example, $S2 = (T, H)$ happens before $S5 = (SBP, L)$, therefore, $R_{2,5} = b$.

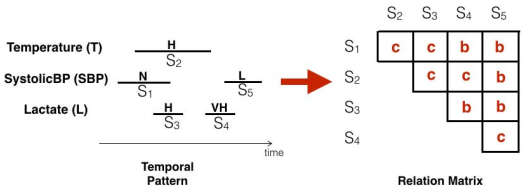


Fig. 3: A temporal pattern P with 5 states $\langle S_1 = (SBP, N), S_2 = (T, H), S_3 = (L, H), S_4 = (L, VH), S_5 = (SBP, L) \rangle$ and temporal relations presented by half matrix R

We define an MSS z_i *contains* a temporal n-pattern P, denoted as $P \in z_i$, if: 1) z_i contains all k states of P, and 2) all temporal relations of P are satisfied in z_i. In step 2, we will find the RTPs from MSSs in Z.

B. Step 2. RTP Mining

We selected the RTP mining algorithm proposed by Batal et al. [15] in this work, mainly because RTP takes the recency of the patterns into consideration by defining a maximum gap parameter g, and secondly, it is an efficient mining algorithm which prevents generating incoherent patterns. In following, we first give the definition of Recent Temporal Pattern (RTP) and then briefly explain how it is applied to our work.

Recent Temporal Pattern: First, we define Recent State Interval between a state interval $E = (F, V, s, e)$ and a MSS z_i. Precisely, we call a State Interval $E = (F, V, s, e)$ is a *Recent State Interval* of z_i:

1) If E is the last state interval for feature F; that is, for all $E' = (F, V', s', e')$, we have $E'.e \leq E.e$; or
2) if E is less than g time units away from the end time of the last state interval: $z_i.end$; that is, $z_i.end - E.e \leq g$.

Further, given an MSS z_i, a temporal pattern $P = (\langle S_1, ..., S_n \rangle, R)$, and a maximum gap parameter g, we say P is a RTP in z_i, denoted as $R_g(P, z_i)$, if *all* of the following conditions hold:

1) z_i contains P: $P \in z_i$; and
2) $S_n = (F_n, V_n)$ matches a recent state interval in z_i; and
3) Any consecutive pair of states in P should map to state intervals that are less than g time units away from each other. That is, each pair of temporal sequences should not be g time units apart.

In short, g is the restrictive parameter which forces the patterns to be close to the end of the sequence z_i and the consecutive states to be close to each other.

Mining Algorithm: For an MSS z_i, we have its corresponding label y_i: whether the outcome of the sequence is septic shock 1 or non-septic shock 0; thus we denote it as $\langle z_i, y_i \rangle$, $y_i \in \{0, 1\}$. By combining all the z_i with the same y_i, we will have two sets of labeled MSSes: $Z_1 = \{z_i : y_i = 1\}$ for all septic shock sequences and $Z_0 = \{z_j : y_i = 0\}$ for all non-septic shock. Given Z_1, the mining algorithm applies a level-wise search to find frequent RTPs. More specifically, it first starts with all frequent 1-RTPs, and then extends the patterns by adding a new state one at a time. That is, at each level k, the algorithm

finds frequent $(k+1)$-RTPs by extending k-RTPs through the following two steps:

1) **Backward candidate generation**: Extending k-patterns by adding a new state to the beginning of the state sequence, to obtain $(k+1)$-pattern candidates.
2) **Counting phase**: Counting the cardinality of occurrence of each candidate and removing the ones which do not meet the minimum support threshold σ.

These two steps repeat until no more $(k+1)$-pattern can be found. Next, we will describe each steps in details.

Backward candidate generation: in order to generate $(k+1)$-pattern candidates from a k-pattern $P = (\langle S_1, ..., S_k \rangle, R)$, we extend the state sequence by adding a new frequent state, S_{new}, to the beginning of the sequence and thus it becomes $P^{'} = (\langle \mathbf{S_{new}}, S_1, ..., S_k \rangle, \mathbf{R}^{'})$. Then we specify the new relations $R^{'}$ between S_{new} and the original k states. For each pair of states, there are two possible temporal relations (*before* or *co-occur*). However, the following two criteria restrict the relations:

- Two state intervals of the same temporal feature cannot co-occur. That is, if $S_{new}.F = S_i.F$ for $i \in \{1, ..., k\}$, then $R'_{new,i} \neq c$.
- Since the state sequence in pattern P is sorted by the start time of the states, once a relation becomes *before*: $R'_{new,i} = b$ for any $i \in \{1, ..., k\}$, all the following relations have to be *before*: $R'_{new,j} = b$ for $j \in \{i+1, ..., k\}$.

Fig. 4 shows an example of generating 3-patterns out of a single 2-RTP P, by appending a new state to the beginning of its states sequence, which results in generating 3 different coherent patterns.

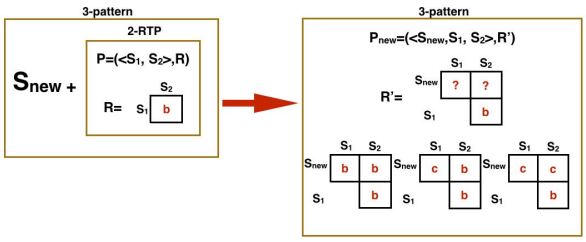


Fig. 4: An example of generating 3-patterns out of a single 2-RTP, by appending a new state.

Counting phase: each of the generated $(k+1)$-patterns are tested for the cardinality of their occurrence as recent temporal pattern among Z_1. If the candidate did not meet the minimum support threshold σ, then it would be removed from the list of $(k+1)$-RTPs.

The same procedure is carried out for Z_0. Note that we explored different σ and g parameters for Z_1 and Z_0. Finally, we combine all the frequent RTPs found into a final set of RTPs as Ω.

C. Step 3. Transform into Binary Matrix.

We transform each MSS $z_i \in Z$ into a binary vector v_i of size $|\Omega|$, such that each 0 and 1 indicates whether the pattern $P_j \in \Omega$ is a recent temporal pattern in Z_i or not. This will result in a binary matrix of size $N \times |\Omega|$, which represents our original dataset.

D. Step 4. Learning the SVM Classifier

Once the binary matrix is built, we apply SVM classifier to perform the target task. Additionally, we explored other classifiers such as logistic regression, decision trees, MLP and SVM, which the latter gave us the best performance and thus is reported in the following.

IV. Experimental Setup

A. Dataset

Our original dataset consists of anonymized clinical multivariate time series data extracted from the EHR system at Christiana Care Health System, from July 2013 to December 2015. Each data point is a visit/episode constituted by a series of events. In total, there are 12,980 visits and 1,446,225 events. Each event consists of critical attributes/features from the following categories: vital signs, lab results, location type code and description.

The *study population* are patients with *suspected infection* as identified by the presence of any type of antibiotic, antiviral, or anti-fungal administration, or a positive test result of Point of Care Rapid (PCR). Note that the study population, the aforementioned rules for identifying suspected infection, and labeling criteria for sepsis progressive stages including septic shock in the following, were all determined by the two leading clinicians with extensive experience on this subject from Mayo Clinic and Christiana Care Health System.

When applying RTP, we used the following features:

- 7 Vital signs: Systolic Blood Pressure, Diastolic Blood Pressure, Heart Rate, Respiratory Rate, Temperature, Oxygen Saturation, Mean Arterial Pressure.
- 10 Lab results: Blood Urea Nitrogen, Procalcitonin, White Blood Cell, Bands, Lactate, Platelet, Creatinine, BiliRubin, CReactiveProtein, Sedimentation Rate.
- Location type: (emergency department, nurse, step down, and intensive care unit), code and description.
- Progressive stages labels prior to septic shock: infection, inflammation, and organ dysfunction; all identified by rules designed by the two clinician experts.

Ground Truth Labeling for Septic Shock: Supervised models depend heavily on the accurate label of the training dataset. However, acquiring the true label (i.e., septic shock and non-septic shock) can be challenging. Diagnosis codes, such as International Classification of Diseases, Ninth Revision (ICD-9), are widely used. However, solely relying on ICD-9 can be problematic as it has been proven to have limited reliability due to the fact that its coding practice is used mainly for administrative and billing purposes. Indeed, it has been widely argued that ICD-9 codes cannot be used for establishing reliable gold standards for various clinical conditions [47]. More importantly, ICD-9 cannot tell when septic shock occurs, which is essential for our task. On the basis of the Third International Consensus Definitions for

Sepsis and Septic Shock [48], our domain experts identified septic shock as having received vasopressor(s) or having had persistent hypotension (i.e.,systolic blood pressure less than 90 mmHg or mean arterial pressure less than 65 mmHg for more than 1 hour).

When applying both ICD-9 and our clinical rules, we identified 1,869 shock positive visits and 23,901 negative visits. Given the imbalanced ratio of positive and negative shock visits, we further conducted a stratified random sampling on shock negative visits while keeping the same underlying distribution of age, gender, ethnicity, length of stay and the number of records in both positive and negative visits. As a result, the final dataset has 3,738 visits (1,869 positives and 1,869 negatives) and 145,421 events consisting of the following two groups:

- The **Cases** are clinical records of 1,869 shock positive visits from the beginning of the visit up to the first *diagnosed* onset of septic shock.
- The **Controls** are the records of 1,869 sampled non-septic shock patients. In order to balance the length of sequences, we cut the controls' sequences to have the same distribution of length as Cases.

A protracted and iterative process was performed for more than two years to transform our raw EHR data into clean and labeled version that we used in this study. This process involves both academicians and healthcare experts. First, the hospital statisticians strictly followed their predefined protocols and guidelines to apply data-merging and pre-processing. Once we gained access to the data, a group of students and faculty also explored the data extensively. In a repeated process, for any inconsistency and concerns in data we directly communicated with hospital statisticians to resolve the issue. In short, we believe our data are properly cleaned and pre-processed by the EHR providers and the human error is minimized.

B. Two Different Early Diagnosis Tasks

We explored two different early diagnosis tasks: the visit level early diagnosis (Left Aligned) and the event level early prediction (Right Aligned).

For the visit level early diagnosis task, we are given the *first up to* n hours of a patient's EHRs and our goal is to predict whether the patient will develop septic shock at any subsequent point during the visit. To conduct this task, we left aligned all patients' EHR sequences by their start times. Our training data includes the EHRs from beginning to the first up[1] to n hours of the visits (see Fig. 5). Our *observation window* here is the first n-hour window.

For the event level early prediction task, we are given all of a patient's EHRs until n hours before the septic shock onset (cases group) or end of the sequences (control group) and our task is to predict whether or not the patient will develop septic shock exactly n hours later. To conduct this task, we *right aligned* all the cases sequences by septic shock onset and all controls by the end of their sequences and include all the EHRs until n hours before the end of sequences (see Fig. 6). Our *prediction window* here is n-hour window before the onset of septic shock or end of sequence.

Generally speaking, the event level early prediction (whether a patient will be in septic shock state exactly n hour later) is more precise than the visit left-aligned task (whether a patient will be in septic shock state at any subsequent points), thus the right-aligned event level early prediction task is more challenging than the left-aligned visit level early diagnosis task.

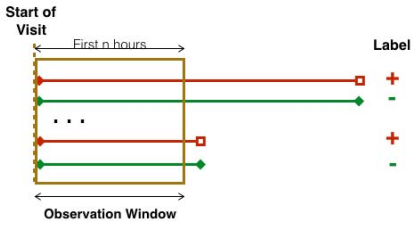


Fig. 5: Visit Level Early Diagnosis (Left Aligned) Setting

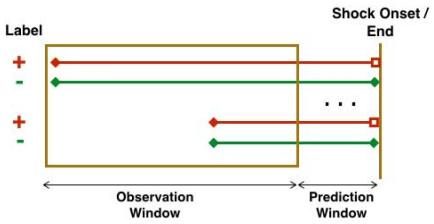


Fig. 6: Event Level Early Prediction (Right Aligned) Setting

C. Two Temporal Abstraction Settings: Entire vs. Truncated

Prior research on applying RTPs for diabetes and heart failure prediction all used *entire* EHR sequences to extract meaningful temporal patterns in the TA step. However, when applying RTP to progressive disease such as septic shock for early prediction, we have the choices of either using the *entire* training sequences or using the *truncated* sequences. Entire means benefiting from the whole sequences of septic shock and non-septic shock visits for pattern extraction, while truncated is referring to adoption of the part of sequence included in observation window to find RTPs. The advantage of using *entire* sequences is that the longer the sequences, the more temporal patterns can be considered and discovered, while the advantage of using *truncated* sequences is that the training data used in TA and RTP mining stages are the same as the testing data for the classification task and thus, the discovered patterns are more likely to emerge and be representative for the early diagnosis task. Thus, we explored RTPs using both the *entire* and the *truncated* sequences for temporal pattern abstraction and named them RTP_{Entire} and RTP_{Trunc} respectively.

[1]If a patient's EHRs are less than n hours, we will include all of them except the last event: septic shock or non-septic shock

D. Experiment Setup

To evaluate the RTP framework for early diagnosis of septic shock, we conducted a series of experiments to test its effectiveness on the two different early diagnosis tasks: left-aligned visit level vs. right-aligned event level. For each task, we used grid search to investigate the optimum value for the maximum gap (g) and minimum support (σ) parameters. The optimum values were: $g = 12$ hours, $\sigma_0 = 0.16$, and $\sigma_1 = 0.14$ for the left-aligned visit level early diagnosis and $g = 14$ hours, $\sigma_0 = 0.2$, and $\sigma_1 = 0.14$ for the right-aligned event level early prediction. For each task, we explored the two RTP settings: RTP_{Entire} and RTP_{Trunc}, and compared them against three categories of baseline methods: *Last Value* from the original paper, the six *Classic ML methods*, and *LSTM*.

- ***Last-Value (LV) based***: This model uses the last measurement of each attribute as the input to a SVM classifier. Similar to the two RTP based methods, we explored both LV_{Entire} and LV_{Trunc}. For the former, we used each attribute's last measurement of entire sequence in the training data to learn the parameters for our SVM classifier while for the LV_{Trunc}, we truncated all the sequences in the training dataset in the same fashion as the testing dataset and then applied the original Last Value approach on the truncated training dataset. For the SVM classifier, we explored different kernel functions, and found that an RBF kernel gave the best results.
- ***Six Classic Machine Learning:*** We explored Logistic Regression, CART Decision Tree (with "Gini" impurity measure), Gaussian Naive Bayes, Random Forest (with max_depth=4), Multi-layer Perceptron (one hidden layer with four hidden units), and SVM (with RBF kernel). Since these classic machine learning baselines do not handle time sequence directly, thus motivated by previous literature [42], the mean, max, min, median and standard deviation of different attributes were calculated as features for numeric variables. For categorical variables, we counted how many times the variables were collected and how many times a given value was obtained.
- ***LSTM:*** This is a variation of Recurrent Neural Network (RNN) that prevents the vanishing gradient problem among other forms of RNN [13]. Recently, this model gained popularity in biomedical domain, since it can effectively model temporal data and capture long range dependencies in sequences. LSTM has a chain-like structure, which enables the information to flow among different blocks at different time points. Each block of the LSTM consists of a memory cell state and three gates: Forget gate, Input gate, and Output gate. These three gates interact with each other to control the flow of information. More specifically, the Forget gate determines what information from the previous memory cell state is expired and should be removed; the Input gate selects information from the candidate memory cell state to update the cell state; the Output gate filters the information from the memory cell so that the model only considers information relevant to the prediction task. Therefore, the memory cell plays a crucial role in memorizing previous experiences. In our task, the input is multivariate temporal sequence of patients, and output from the last step is used to make prediction. We implemented LSTM in Keras with Tensorflow as the back-end engine, and we used one hidden layer with 100 hidden neurons and maximum length as the longest sequence in data. The whole multivariate time series from EHR data were fed to this model as input data, with mean imputation for missing values. We applied 5-fold nested cross-validation in order to tune the parameters of the model, such as optimizer, initializer, number of epochs, and number of batches.

E. Evaluation Metrics

F1 Score, AUC, accuracy, recall and precision were computed to evaluate our models. Accuracy represents the proportion of patients whose labels were correctly identified. Recall tells us what proportion of patients that actually had septic shock were correctly diagnosed by the model as having septic shock. Precision tells us what proportion of patients who were diagnosed as having septic shock actually had septic shock. F1 Score is the harmonic mean of Precision and Recall that sets their trade-off. AUC calculates the tradeoff between recall and specificity. Therefore, in the following we will mainly use F1-score and AUC to compare different models. All models were evaluated using 5-fold cross validation.

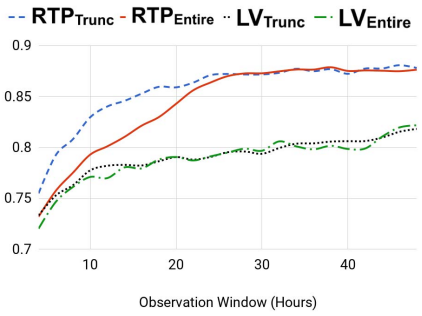
V. Results and Discussion

In the following sections, we present our results in three parts, First, motivated by the original study, we compare RTP against the original baseline approach: LV, to see whether the same results stand for progressive disease: septic shock. Additionally, we compare RTP and LV on both settings using the entire sequences and the truncated sequences. Second, we will further explore the robustness of the RTP by comparing it against the six classic machine learning methods and the state of the art approach: LSTM. Finally, we will shed light on the extracted interpretative and meaningful temporal patterns discovered by RTP mining algorithm.

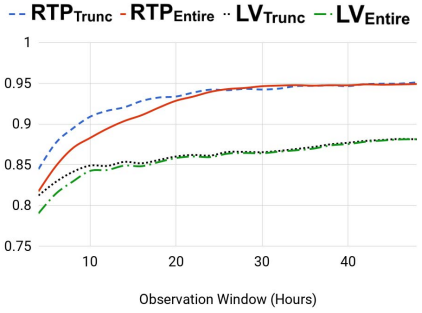
A. RTP vs. Last-Value (LV) Baseline

We compared the performance of RTP vs. LV on two different early diagnosis tasks and for each task, we examined both methods using two different portion of data for pattern extraction or as last measurement: *entire* and *truncated*.

For left-aligned visit level early diagnosis, we vary the duration of the observations from the first 4 hours up to the first 48 hours incrementing by 1 hour. Therefore, the longer the duration, the better performance is expected. Fig. 7 shows the F-measure and AUC performance of the two RTP-based models with the two LV baselines. As we can see, the two RTPs consistently outperform the two LV methods. Moreover, between the two RTP-based models, RTP_{Trunc} performs better than RTP_{Entire} up to 26 hours observation window

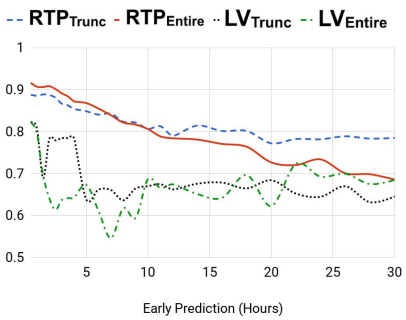


(a) F-measure performance

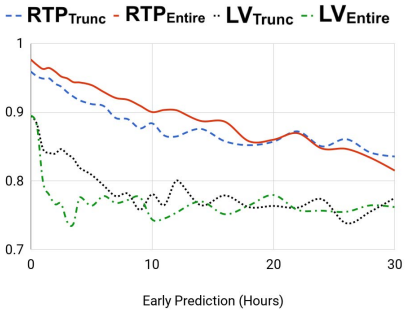


(b) Area under ROC performance

Fig. 7: *Visit diagnosis* performance of RTP-based and last-value models, with two different feature extraction data: Entire and Truncated.



(a) F-measure performance



(b) Area under ROC performance

Fig. 8: *Early prediction* performance of RTP-based and last-value models, with two different feature extraction data: Entire and Truncated.

and after that, the two perform very closely. This is because most of the visit lengths are less than 26 hours and the two training datasets become very similar when the observation window is expanded more than 26 hours.

For the right-aligned event-level early prediction task, we vary the duration of the prediction window from zero up to 30 hours before the onset of septic shock incrementing by 1 hour. As we move the duration of the observations further away from onset or the end of the sequences, the number of Cases and Controls decrease. That is, the further away from the moment of septic shock onset, the more challenging the early prediction task is and thus the worse performance is expected. Fig. 8 shows F-measure and AUC performance of the two RTP-based classifiers vs. the two LV-based baselines. As expected, the performance drops as we move observation window further away from the outcome point, while the performance of the two LVs drop more rapidly comparing to the two RTPs. Fig. 8(a) shows that when the prediction window is < 8 hours, RTP_{Entire} has slightly better performance than RTP_{Trunc} on F-measure. However, when the prediction window is moved ≥ 8 hours, RTP_{Trunc} consistently outperforms RTP_{Entire} and the two LV baselines.More importantly, RTP_{Trunc}'s performance stays stable and acceptable (slightly below 0.8 around 30h). Fig. 8(b) shows their AUC results; first, the two RTP models significantly outperform the two LV methods; second, while RTP_{Entire} has better AUC performance than RTP_{Trunc} in the left part of the figure, the performance of both RTP_{Trunc} and RTP_{Entire} are above 0.8 even when prediction window reaches 30 hours. When combining F-measure and AUC performance, our results suggest that RTP_{Trunc} is a better model in that as the early prediction task becomes more challenging (prediction window is expanded), the performance of RTP_{Trunc} stays reasonable. This is probably because by using the truncated training data for pattern extraction in RTP_{Trunc}, it is more likely to capture the recent temporal patterns that are not only predictive of septic shock onset but also more likely to be observed in the testing dataset.

B. RTP vs. Classic ML & LSTM

Table I shows different performance measures of six classic machine learning methods, LSTM, and the two RTP-based approaches for the task of left-aligned visit-level early diagnosis using the first up to 8 hours observations. The top section of Table I shows that SVM outperforms the other five classic classifiers across all the five measures and LSTM outperforms all six classic ML classifiers across all measures except precision. Our results is consistent with previous work that deep learning models such as LSTM is indeed more effective than the classic ML approaches. Despite all these, Table I shows that RTP_{Entire} performs worse than LSTM

TABLE I: Visit diagnosis performances for *eight* hours observation window.

	Model	Accuracy	Precision	Recall	F-measure	AUC
	Quantitative TA + Classic ML					
1	Logistic Regression	0.751	0.771	0.714	0.741	0.823
2	Decision Tree	0.724	0.727	0.715	0.721	0.725
3	Gaussian Naive Bayes	0.701	0.830	0.505	0.628	0.817
4	Random Forest	0.777	0.817	0.714	0.762	0.857
5	MLP	0.772	0.787	0.745	0.765	0.848
6	SVM	0.806	**0.836**	0.761	0.797	0.878
	Deep Learning					
7	LSTM	0.808	0.821	0.788	0.804	0.892
	RTP-based Classification					
8	RTP_Trunc	**0.813**	0.830	0.787	**0.808**	**0.895**
9	RTP_Entire	0.728	0.660	**0.938**	0.775	0.871

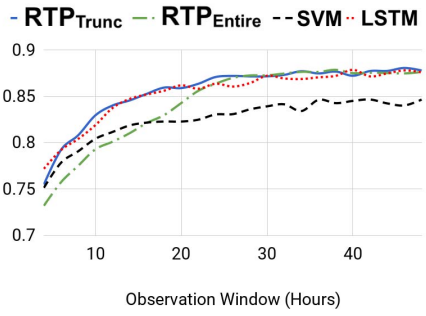
TABLE II: Early prediction performance for *four* hours prediction window.

	Model	Accuracy	Precision	Recall	F-measure	AUC
	Quantitative TA + Classic ML					
1	Logistic Regression	0.802	0.802	0.802	0.802	0.846
2	Decision Tree	0.752	0.741	0.774	0.757	0.731
3	Gaussian Naive Bayes	0.657	0.778	0.439	0.562	0.722
4	Random Forest	0.799	0.822	0.762	0.791	0.874
5	MLP	0.788	0.769	0.823	0.795	0.844
6	SVM	0.791	0.763	0.842	0.801	0.861
	Deep Learning					
7	LSTM	0.839	0.829	0.855	0.842	0.916
	RTP-based Classification					
8	RTP_Trunc	0.847	0.840	**0.858**	0.849	0.917
9	RTP_Entire	**0.875**	**0.915**	0.826	**0.868**	**0.943**

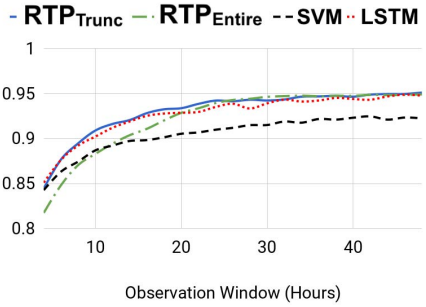
and even some of the classic machine learning methods across all measures except on recall, it reaches the significant results of 0.938; RTP_{Trunc}, on the other hand, outperforms LSTM on all measures except on recall, it has slighly worse performance than LSTM: 0.787 vs. 0.788. RTP_{Trunc} also outperforms all classic classifiers across all measures except on precision, it has slightly worse performance than SVM: 0.830 vs. 0.836. In short, it seems that RTP_{Trunc} has the best overall performance when using the first up to 8 hours EHRs.

Table II shows the same comparison for the task of right-aligned event-level early prediction for 4 hours before the onset of the septic shock. The top section of Table II shows that while there is no clear winner among the six classic classifiers, LSTM is again more effective than the six classifiers on this task in that LSTM significantly outperforms the latter across all the five measures. Finally, Table II shows that RTP_{Trunc} outperforms LSTM on all measures and the former also outperforms all the six classic classifiers across all measures. Similarly, RTP_{Entire} performs significantly better than LSTM across all measures except on recall and RTP_{Entire} also outperforms all six classic machine learning methods across all measures except on recall. Among all the models, RTP_{Entire} performed best on four out of five measures and the only exception is on recall but its value is still reasonable. Therefore, it seems that RTP_{Entire} is a better model when using all EHRs up to 4 hour before the onset of septic shock.

Based on the results from Table I and II, considering temporal aspect of EHR data is the key factor for precise prediction, since both LSTM and RTP constantly outperform the



(a) F-measure performance



(b) Area under ROC performance

Fig. 9: Early prediction performance

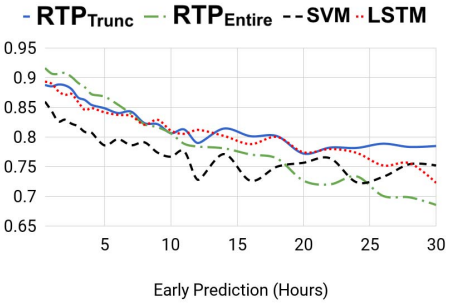
classic atemporal models. Therefore, to examine the strength of the RTP-based models for various observation windows, we compared it against LSTM and the most powerful classic ML model, SVM. Fig. 9 compares F-measure and AUC performance of SVM, RTP_{Trunc}, RTP_{Entire}, and LSTM up to 48 hours prediction window in left aligned setting. As we can see, both LSTM and RTP-based models outperform SVM at all times. Also, RTP_{Trunc} can perform highly comparable and close to LSTM, while having higher AUC for almost all observation windows. Fig. 10 presents early prediction performance of the same methods up to 30 hours prediction window in right aligned setting. Similarly, in this setting all temporal models outperform SVM. LSTM and RTP_{Trunc} perform highly analogous, while RTP_{Entire} outperforms both in AUC measure up to 20 hours prediction window. Indeed, RTP-based models outperform LSTM almost at all times, which reveals the fact that some latent patterns are captured by RTP_{Entire} that are highly predictive of future events.

C. Knowledge Discovery

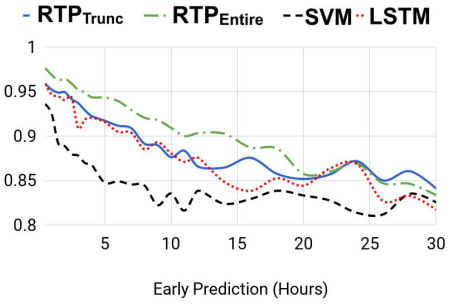
One substantial advantage of pattern-based classification over deep learning methods, particularly for clinical event prediction, is *interpretability* of the patterns. In such models, the patterns need to be representative of the original time series data while predictive of the future events. In this study, most of the patterns extracted using RTP mining are in accordance with the medical diagnosis guidelines and some of them reveal latent patterns towards progression of septic shock.

TABLE III: Recent temporal patterns extracted right before the onset of septic shock, with $\sigma_0 = 16\%$, $\sigma_1 = 14\%$ and $g = 12$h.

RTP	If	Then	Support
P_1	(Location,ED) *c* ((Inflammation,0) *b* ((Inflammation,1) *c* (OrganFailure,1)))	Shock	0.367
P_2	(Location,ED) *c* ((Inflammation,0) *b* (PulseOx,VH))	Shock	0.274
P_3	(Location,ED) *c* ((Inflammation,0) *b* ((RespiratoryRate,VH) *c* (Inflammation,1))	Shock	0.192
P_4	(Location,ED) *c* ((Inflammation,0) *b* ((BUN,N) *c* (Creatinine,N) *c* (Inflammation,1))	Shock	0.161
P_5	((Location,ED) *c* ((Inflammation,0) *b* ((Inflammation,1) *c* (Creatinine,N))) *b* (Location,NURSE)	Non-shock	0.186
P_6	(Platelet,N) *c* (WBC,N)	Non-shock	0.327
P_7	(RespiratoryRate,L) *b* (RespiratoryRate,N)	Non-shock	0.201



(a) F-measure performance



(b) Area under ROC performance

Fig. 10: Visit diagnosis performance

Table III presents a number of interesting patterns and their corresponding support among the training group. P_1 describes the most pronounced rule in septic shock progression in emergency department, when inflammation is not observable at first, then in less than 12 hours inflammation and organ failure symptoms develop. This pattern is observable among 36.7% of patients right before the onset of septic shock. Most of the frequent patterns among shock patients are happening in emergency department. P_2 and P_3 illustrate when patient is in emergency department with no sign of inflammation in the beginning. In the meantime, if oxygen saturation or respiratory rate becomes very high after 12 hours, the patient is prone to develop septic shock in the next 12 hours. P_4 indicates the case where Bun and Creatinine are normal while inflammation is developing. This reveals the fact that blood sample was taken from those patients, however, these two factors were frequently normal among them.

P_6, which is a 5-pattern, describes a situation that patient is in emergency department and even though the inflammation symptoms appear, since the test results (like Creatinine) are normal and stable, patient is moved to nursing unit. Indeed, glancing over the extracted patterns among non-shock group, we can see that most of the patients from emergency department end up in a stable state and are eventually moved to nursing unit. Additionally, most of the frequent patterns show normal vital signs and lab values, or changing from an abnormal state to normal states, like patterns P_6 and P_7 in the Table III.

VI. Conclusion

Early prediction of septic shock is a challenging task due to subtle progression and unspecific symptoms associated with this fast-progressive disease. In this study, we integrated EHRs, clinical expertise, and various machine learning algorithms to build a classifier that is able to predict septic shock before onset with high accuracy, and to provide valuable insight for clinicians. We employed an RTP-based classification framework and compared it with various baselines including classic and deep learning models, in two different settings of visit-level early diagnosis and event-level early prediction. The results suggest that the RTP framework consistently outperforms the atemporal classic baselines in both settings at all hours before the onset. Moreover, the RTP_{Trunc} model can perform comparable to LSTM in left aligned setting. However, in right aligned setting, RTP_{Entire} outperforms LSTM up to 20 hours early prediction. The AUC early prediction performance of this model is above 85% almost at all times.

As future work, we plan to apply pre-clustering of patients to discover more definitive patterns and obtain more advantageous knowledge discovery as well as improved prediction performance. Additionally, this work will be applied to larger EHR dataset from Mayo Clinic and public dataset MIMIC-III, along with integration of more informative features such as intervention/medication and demographic features to develop more robust models.

Acknowledgements: This research was supported by the NSF Grant #1522107, S.E.P.S.I.S. (Sepsis Early Prediction Support Implementation System).

References

[1] C.M. Torio and R.M. Andrews. National inpatient hospital costs: the most expensive conditions by payer, 2013., 2016.

[2] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.

[3] G.S Martin, D.M Mannino, et al. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 2003.

[4] R.P. Dellinger, M.M Levy, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive care medicine*, 2008.

[5] A. Kumar, D. Roberts, 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), June 2006.

[6] V. Coba, M. Whitmill, et al. Resuscitation bundle compliance in severe sepsis and septic shock: improves survival, is better late than never. *Journal of intensive care medicine*, 2011.

[7] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

[8] J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and LG Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710, 1996.

[9] Nathan I Shapiro, Richard E Wolfe, Richard B Moore, Eric Smith, Elizabeth Burdick, and David W Bates. Mortality in emergency department sepsis (meds) score: a prospectively derived and validated clinical prediction rule. *Critical care medicine*, 31(3):670–675, 2003.

[10] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[11] Jordi Rello, Alejandro Rodriguez, Thiago Lisboa, Miguel Gallego, Manel Lujan, and Richard Wunderink. Piro score for community-acquired pneumonia: a new prediction rule for assessment of severity in intensive care unit patients with community-acquired pneumonia. *Critical care medicine*, 37(2):456–462, 2009.

[12] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[15] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–288. ACM, 2012.

[16] Kalia Orphanou, Arianna Dagliati, Lucia Sacchi, Athena Stassopoulou, Elpida Keravnou, and Riccardo Bellazzi. Combining naive bayes classifiers with temporal association rules for coronary heart disease diagnosis. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 81–92. IEEE, 2016.

[17] Yuval Shahar. A framework for knowledge-based temporal abstraction. *Artificial intelligence*, 90(1-2):79–133, 1997.

[18] Lucia Sacchi, Cristiana Larizza, Carlo Combi, and Riccardo Bellazzi. Data mining with temporal abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2):217–247, 2007.

[19] Tudor Toma, Ameen Abu-Hanna, and Robert-Jan Bosman. Discovery and inclusion of sofa score episodes in mortality prediction. *Journal of Biomedical Informatics*, 40(6):649–660, 2007.

[20] Shin-Yi Wu and Yen-Liang Chen. Mining nonambiguous temporal patterns for interval-based events. *IEEE transactions on knowledge and data engineering*, 19(6), 2007.

[21] Hong Cheng, Xifeng Yan, Jiawei Han, and S Yu Philip. Direct discriminative pattern mining for effective classification. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 169–178. IEEE, 2008.

[22] Robert Moskovitch and Yuval Shahar. Fast time intervals mining using the transitivity of temporal relations. *Knowledge and Information Systems*, 42(1):21–48, 2015.

[23] Dmitriy Fradkin and Fabian Mörchen. Mining sequential patterns for classification. *Knowledge and Information Systems*, 45(3):731–749, 2015.

[24] Iyad Batal, Hamed Valizadegan, Gregory F Cooper, and Milos Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):63, 2013.

[25] Dhaval Patel, Wynne Hsu, and Mong Li Lee. Mining relationships among interval-based events for classification. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 393–404. ACM, 2008.

[26] Gessé Dafé, Adriano Veloso, Mohammed Zaki, and Wagner Meira. Learning sequential classifiers from long and noisy discrete-event sequences efficiently. *Data Mining and Knowledge Discovery*, 29(6):1685–1708, 2015.

[27] Robert Moskovitch, Colin Walsh, George Hripcsak, and NP Tatonetti. Prediction of biomedical events via time intervals mining. In *NYC, USA: ACM KDD Workshop on Connected Health in Big Data Era*, 2014.

[28] Robert Moskovitch and Yuval Shahar. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA annual symposium proceedings*, volume 2009, page 452. American Medical Informatics Association, 2009.

[29] Robert Moskovitch and Yuval Shahar. Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*, 29(4):871–913, 2015.

[30] Shameek Ghosh, Jinyan Li, Longbing Cao, and Kotagiri Ramamohanarao. Septic shock prediction for icu patients via coupled hmm walking on sequential contrast patterns. *Journal of biomedical informatics*, 66:19–31, 2017.

[31] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

[32] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[33] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2016.

[34] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[35] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[36] Femida Gwadry-Sridhar, Benoit Lewden, Selam Mequanint, and Michael Bauer. Comparison of analytic approaches for determining variables-a case study in predicting the likelihood of sepsis. In *HEALTHINF*, pages 90–96, 2009.

[37] Steven W Thiel, Jamie M Rosini, William Shannon, Joshua A Doherty, Scott T Micek, and Marin H Kollef. Early prediction of septic shock in hospitalized patients. *Journal of hospital medicine*, 5(1):19–25, 2010.

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

[39] Susana M Vieira, Luís F Mendonça, Gonçalo J Farinha, and João MC Sousa. Modified binary pso for feature selection using svm applied to mortality prediction of septic patients. *Applied Soft Computing*, 13(8):3494–3504, 2013.

[40] Jürgen Paetz. Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. *Artificial Intelligence in Medicine*, 28(2):207–230, 2003.

[41] Peter Haug and Jeffrey Ferraro. Using a semi-automated modeling environment to construct a bayesian, sepsis diagnostic system. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 571–578. ACM, 2016.

[42] Dewang Shavdia. *Septic shock: Providing early warnings through multivariate logistic regression models*. PhD thesis, Massachusetts Institute of Technology, 2007.

[43] Caleb Hug. *Detecting hazardous intensive care patient episodes using real-time mortality models*. PhD thesis, 2009.

[44] Marta Carrara, Giuseppe Baselli, and Manuela Ferrario. Mortality prediction in septic shock patients: Towards new personalized models in critical care. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 2792–2795. IEEE, 2015.

[45] Yuan Zhang, Chen Lin, Min Chi, Julie Ivy, Muge Capan, and Jeanne Huddleston. Lstm for septic shock: Adding unreliable labels to reliable predictions. pages 1233–1242, 12 2017.

[46] James F Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984.

[47] K.K. Giuliano. Physiological monitoring for critically illpatients: testing a predictive model for the early detection of sepsis. *AJCC*, 16(2), March 2007.

[48] M. Singer, C.S. Deutschman, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *AMA*, 315(8), February 2016.