Time-aware Subgroup Matrix Decomposition: Imputing Missing Data Using Forecasting Events

Xi Yang, Yuan Zhang, Min Chi
Department of Computer Science, College of Engineering
North Carolina State University, Raleigh, NC, USA
Email: {yxi2, yzhang93, mchi}@ncsu.edu

***Abstract*—Deep neural network models, especially Long Short Term Memory (LSTM), have shown great success in analyzing Electronic Health Records (EHRs) due to their ability to capture temporal dependencies in time series data. When applying the deep learning models to EHRs, we are generally confronted with two major challenges: high rate of *missingness* and *time irregularity*. Motivated by the original PACIFIER framework which utilized matrix decomposition for data imputation, we applied and further extended it by including three components: forecasting future events, a time-aware mechanism, and a subgroup basis approach. We evaluated the proposed framework with real-world EHRs which consists of 52,919 visits and 4,224,567 events on a task of early prediction of septic shock. We compared our work against multiple baselines including the original PACIFIER using both LSTM and Time-aware LSTM (T-LSTM). Experimental results showed that our proposed framework significantly outperformed all competitive baseline approaches. More importantly, the extracted interpretative latent patterns from subgroups could shed some lights for clinicians to discover the progression of septic shock patients.**

***Keywords*-imputation; forecasting future events; irregular interval; time-aware; subgroup; septic shock early prediction**

I. Introduction

Electronic Health Records (EHRs) are large-scale systematic collections of sequential data which include both static and dynamic information for each patient's visit [1]. The *static* information, such as age and sex, is often collected once per visit and remains unchanged for the whole duration of the visit; whereas *dynamic* information is generally collected with different frequencies. For example, the body temperature is often measured several times a day, while the white blood cells are measured every other day. As a result of merging such irregular dynamic information, real-world EHRs are often plagued by missing data problem. To tackle this issue, various imputation strategies have been explored. Some common approaches include mean- or median-filling, carrying forward, hot-deck, resampling [2], multiple imputation [3], and so on. Recently, Lipton et al. indicated using missing indicators [4] is highly effective for handling temporal missing data [5]. Kim et al. proposed a bio-inspired approach named Temporal Belief Memory for handling the missingness in sequential data with irregular intervals [6]. Some other approaches focused on reconstructing missing entries using latent patterns extracted from the original data, such as EM imputation [7], Autoencoder [8], and matrix decomposition based imputation methods including SVDImpute [9], softImpute [10], Individual Basis Approach (IBA) and Shared Basis Approach (SBA) [11]. In this work, we focused on *matrix decomposition based* imputation methods because of their robustness yet interpretability, which are well-suited for EHRs. Specifically, the assumption is that the observed features in EHRs can be mapped to some latent medical condition, thus the matrix decomposition based methods could impute the missing entries by exploring the latent structures on both feature and time dimensions.

In our work, we extended the original PACIFIER framework [11] by forecasting future events, incorporating a time-aware mechanism, and employing a subgroups based approach. For *forecasting future events*, we explored the potentials of applying the extracted latent patterns to forecast future events, and then use these forecasted future events to further improve the effectiveness of our model. The *time-aware mechanism* would take the time irregularity into consideration and capture latent patterns constrained by the irregular time intervals of inputs to handle the missing data. The *subgroups basis approach* would cluster patients into subgroups by static information and learn latent patterns for each subgroup, since it is shown that static information such as comorbidities, age [12], and gender [13] are important predictors for many diseases such as Parkinson's disease. The subgroup basis approach can be considered as a trade-off between either IBA that learns latent patterns for each individual patient or SBA that learned latent patterns for the entire population explored in [11]. We evaluated our extended framework by comparing it to five competitive missing data handling methods: carrying forward, mean imputation, and three matrix decomposition based methods, i.e., SVDImpute [9], IBA [11], and SBA [11]. The experimental results showed that our proposed framework can effectively handle the EHRs with high missing rate, and it outperformed all five baseline methods including IBA and SBA. Consider the three extended components in our framework, it is referred as ***Time-aware subGroup Basis Approach with Forecasted events (TGBA-F)*** hereinafter. Note that the forecasting future events, time-aware mechanism, and subgroups basis approach tackle missing data from three different perspectives in that the best performance is obtained when we combine all three.

In recent years, Recurrent Neural Networks (RNNs) and its variations such as Long Short-Term Memory (LSTM) [14], Gated Recurrent Unit (GRU) [15], and Time-aware LSTM (T-LSTM) [16], have achieved state-of-the-art results in many real-word applications with multivariate temporal data through deep hierarchical feature construction. Moreover, these variations are capable of capturing long-range dependencies in time series data in an effective manner. Missing data imputation has been widely studied in previous RNN-based works [17] and applied for speech recognition and blood-glucose prediction [18]. Recently, researchers tried to handle the missing data problem in RNNs by concatenating missing entries, incorporating a time based decay function, and synchronizing different sampling frequencies [5], [19], [20]. To our best knowledge, no prior work has explored applying matrix decomposition based imputation methods with RNN-based models, such as LSTM and T-LSTM, to temporal data.

Herein, our task is early prediction of septic shock. Sepsis is a life-threatening organ dysfunction caused by deregulated host response to infection [21]. As the most severe stage of sepsis, *septic shock* reaches a mortality rate as high as 50% and the annualized incidence keeps rising [22]. Prior studies have indicated that the early diagnosis and treatment of septic shock can prevent about 80% of sepsis death. Besides, over the first 6 hours after the onset of recurrent or persistent hypotension, every hour delay in antibiotic treatment leads to a 7.6% decrease in survival of septic patients [23]. One major challenge associated with early prediction of sepsis/septic shock is its subtle but fast progression at early stages. Sepsis has a wide range of potential symptoms, and its common indicators such as infection, fast heart rate, high/low body temperature, and low blood pressure [24] are highly likely to progress to other disease. Because of such delicate progression, variables in the before-shock stage may either be measured infrequently or not measured at all. As a result, the duration between two clinical events in EHRs can be long and the missing rate can be very high. For example, for the EHRs utilized in this work, the missing rate is higher than 80% on average and several variables' missing rates are above 99.9%. Thus, missing data handling is a key factor in our early prediction of septic shock.

The remaining parts of this paper are organized as follows. In Section II, related works are reviewed. Section III presents matrix decomposition, the three components in TGBA-F, and two classifiers: LSTM and T-LSTM. In Section IV, we discuss experimental setup and introduce early prediction models and baselines. Section V presents the experimental results. Finally, Section VI concludes the paper.

II. Background

A variety of approaches have been proposed to cope with the missing data in EHRs, including carrying forward, mean imputation, k-nearest neighbors [25], autoencoder [8], and several matrix decomposition based imputation methods, such as SVDImpute [9] and softImpute [10], which are both Singular Vector Decomposition (SVD)-based methods. Closely related to this work, Zhou et al. proposed a matrix decomposition based imputation framework named PACIFIER (PAtient reCord densIFIER) [11]. Their results showed that PACIFIER can not only impute the missing data, but also imply some macro phenotypes through decomposed latent patterns. In their work, the imputed data would be fed into a logistic regression classifier for early prediction.

In recent years, deep neural network models, especially the RNN-based models, have shown great success in analyzing EHRs due to their ability of capturing temporal dependencies in time series data. Although RNNs are theoretically capable of finding the long-term dependencies underlying the temporal data, classical RNNs often cannot effectively capture the long-term dependencies due to the vanishing and exploding gradient problem [26]. As variations of RNNs, LSTM and GRU can overcome these issues by incorporating multiple gating units into RNN structure. The gating mechanism allows for explicit memory delete and update, and controls flow of information in hidden states. In standard LSTM, it is assumed that intervals between consecutive events are uniform. To consider the time irregularity, several previous works have been proposed with RNN-based models [16], [27], [28]. Among them, Time-aware LSTM (T-LSTM) [16] transforms time intervals into weights to adjust the memory passed from the previous memory cell.

Standard imputation methods have been widely used when applying RNN-based models. For instance, in [16], Baytas et al. employed carrying forward for imputation when evaluating the T-LSTM. As a frontier work of the missing data handling in RNN-based models, [17] proposed an RNN structure for both missing inputs and asynchronous data, which randomly initialized missing values and optimized the imputed values by backpropagation. In [18], they demonstrated a modified RNN for missing data handling, which is combined with a linear error model and trained by expectation-maximization technique. The experimental results showed that their method improved the performance in glucose/insulin metabolism prediction with respect to both conventional RNNs and various linear models. Recently, [5] showed the effectiveness of Missing Indicators (MI) with LSTM for a phenotype prediction task using EHRs. In their work, they gave an insight that LSTM could implicitly impute missing values based on its memory. About the same period, a Phased LSTM [20] was proposed to extend LSTM unit by adding a time gate to align asynchronous streams, which allowed the feature learning only when the time gate is open. Besides, [29] incorporated a carrying forward operation for missing data in RNN and LSTM, and then tested it with a clinical variable prediction task. More recently, GRU-D [19] imputed missing values using a modified GRU, regulated by a temporal decay function with trainable weights, and on a wide range of tasks, the authors showed that GRU-D often demonstrated performance comparable to MI. Furthermore, Temporal Belief Memory (TBM) [6] could systematically impute missing values in both forward and backward directions within a reliable time window, and their results showed that TBM outperformed a wide range of baseline methods.

In short, both matrix decomposition based and RNN-based imputation methods have been widely explored in prior work. However, as far as we know, no previous work has investigated on applying matrix decomposition based imputation methods with RNN-based models. On one hand, matrix decomposition based imputation methods can extract meaningful and interpretable latent patterns represented by the mappings from features (symptoms) to latent patterns (medical conditions) from the original data; on the other hand, RNN-based models can capture temporal dependencies in time series data. Therefore, we expect that by combining them two, we can not only learn more interpretable latent patterns, but also achieve better early prediction performance.

Our work in this paper is highly motivated by Zhou et al.'s PACIFIER framework. Both their work and ours are conducted on so-called *event level early prediction task*, that is, to predict whether a patient will develop to a target medical disease τ hours later. To do so, all sequences are *right aligned* by their endpoint and only the truncated EHRs happened τ-hour before the endpoint, referred as observation data, are used for early prediction. Our work differs from the original PACIFIER in the following four aspects. First, we applied the extracted latent patterns to forecast future events. As a result, rather than only using the observation data, we integrated the *forecasted* future events with our observation data for early prediction. To forecast future events, we explored using truncated EHRs or using the entire sequences to induce the latent patterns while original PACIFIER used truncated EHRs only. Second, while original PACIFIER treated all time intervals equally, we proposed a time-aware mechanism for the smoothness regularization when doing matrix decomposition. It is done by using time intervals as weights to control the smoothness of imputed data. Since the events in EHRs were unevenly collected with irregular time intervals, the time-aware constraints enabled the imputed data to reflect the progression of sequence more accurately. Third, we proposed a subgroup basis approach using static information such as age and sex to partition patients into some subgroups and learned latent patterns for each subgroup while original PACIFIER framework investigated on either learning latent patterns for each individual patient (IBA) or learning latent patterns for the entire population (SBA). Last, original PACIFIER framework was evaluated using logistic regression while we evaluated our framework using two state-of-the-art deep learning models: LSTM and T-LSTM.

III. Method

In a nutshell, our framework contains two stages: *data imputation stage* using our proposed Time-aware sub-Group Basis Approach with Forecasted events (TGBA-F) and *classification stage* using LSTM and T-LSTM.

A. Data Imputation Stage

We have $\mathbf{X}^i \in \mathbb{R}^{m \times e^i}$ denoting the sequence of a patient's visit i, with m features and e^i events, and the total number of visits (sequences) is N. $\mathbf{\Omega}^i$ is a location indicator vector for accessing the observable entries in $\mathbf{X}^i$. A projection operator $\mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i)$ is employed to reserve the observable entries and to convert the missing entries to 0, i.e., if $(p,q) \in \mathbf{\Omega}^i$, then $\mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i)_{(p,q)} = \mathbf{X}^i_{(p,q)}$; otherwise, $\mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i)_{(p,q)} = 0$, where $p \in [1,m]$ and $q \in [1,e^i]$. The core idea of matrix decomposition approach is to decompose the original matrix $\mathbf{X}^i$ into $\hat{\mathbf{U}}\mathbf{V}^i$, where $\hat{\mathbf{U}} \in \mathbb{R}^{m \times k}$ and $\mathbf{V}^i \in \mathbb{R}^{k \times e^i}$ indicate the mappings from m features to k latent patterns, and from these latent patterns to e^i events, respectively. The decomposition result $\hat{\mathbf{U}}\mathbf{V}^i$ is expected to keep the same entries as $\mathbf{X}^i$ at the locations of $\mathbf{\Omega}^i$, i.e., $\left\|\mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i) - \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{U}^i\mathbf{V}^i)\right\|_F^2$ is supposed to be minimized.

In the original PACIFIER framework, the optimization procedure is done by introducing an intermediate matrix $\mathbf{S}^i$, which is taken as a delegation for $\mathbf{X}^i$, where $\mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{S}^i) = \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i)$. More specifically, the objective function is:

$$\begin{aligned} \min_{\mathbf{S}^i,\hat{\mathbf{U}},\mathbf{V}^i} \; & \sum_{i=1}^{N} \frac{1}{2e^i} \underbrace{\left\|\mathbf{S}^i - \hat{\mathbf{U}}\mathbf{V}^i\right\|_F^2}_{\mathcal{D}(\mathbf{S}^i,\hat{\mathbf{U}},\mathbf{V}^i)} + \lambda_1 \underbrace{\left\|\hat{\mathbf{U}}\right\|_1}_{\mathcal{R}_1(\hat{\mathbf{U}})} \\ & + \lambda_2 \sum_{i=1}^{N} \frac{1}{2e^i} \underbrace{\left\|\mathbf{V}^i\right\|_F^2}_{\mathcal{R}_2(\mathbf{V}^i)} + \lambda_3 \sum_{i=1}^{N} \frac{1}{2e^i} \underbrace{\left\|\mathbf{V}^i\mathbf{Z}^i\right\|_F^2}_{\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i)} \\ & \text{s.t. } \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{S}^i) = \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^i), \; \hat{\mathbf{U}} \geq 0 \end{aligned} \tag{1}$$

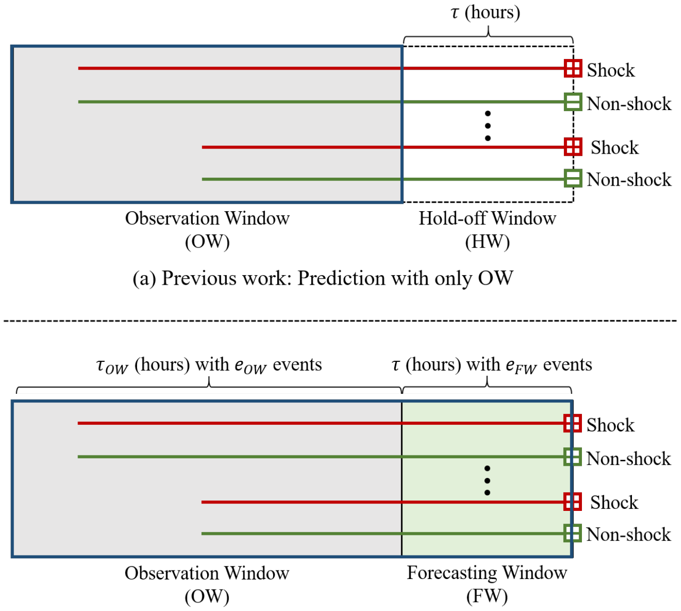
The effects of each term in Eq.(1) are as follows:

- $\mathcal{D}(\mathbf{S}^i,\hat{\mathbf{U}},\mathbf{V}^i)$: *Data fitting term* which ensures the decomposed result $\hat{\mathbf{U}}\mathbf{V}^i$ to be close to the intermediate matrix $\mathbf{S}^i$;
- $\mathcal{R}_1(\hat{\mathbf{U}})$: *Sparseness term* which controls the sparseness of $\hat{\mathbf{U}}$ via a l_1-norm, therefore only the most significant features are involved in mapping to each latent pattern;
- $\mathcal{R}_2(\mathbf{V}^i)$: *Overfitting term* which prevents the decomposition from overfitting through an F-norm. It can circumvent large variances in $\mathbf{V}^i$, thereby controls the model complexity;
- $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i)$: *Smoothness term* which enables temporal latent patterns in $\mathbf{V}^i$ to develop smoothly. Herein, $\mathbf{Z}^i \in \mathbb{R}^{e^i \times (e^i-1)}$ is defined in Eq.(2), so that $\mathbf{V}^i\mathbf{Z}^i$ indicates pairwise differences between consecutive events in $\mathbf{V}^i$ along the trajectory.

$$\mathbf{Z}^i_{(p,q)} = \begin{cases} 1 & \text{If } q = p, \text{ for } p \in [1, e^i - 1], \\ -1 & \text{Elseif } q = p - 1, \text{ for } p \in [2, e^i], \\ 0 & \text{Otherwise.} \end{cases} \tag{2}$$

Since features collected in EHRs are generally positive and latent patterns represented by these features are also supposed to be positive, $\hat{\mathbf{U}}$ is imposed to be larger than 0 in Eqs.(1). To solve for the non-convex objective function in Eqs.(1), a block coordinate descent (BCD) algorithm is used [11]. It follows a turn-taking manner to solve the multiple variables one by one. Specifically, for the three variables, i.e., $\mathbf{S}^i$, $\hat{\mathbf{U}}$ and $\mathbf{V}^i$, in each iteration, one of them is updated with remaining two fixed. Each sub-problem can be solved by the existing optimization solvers. Here, we used a MALSAR package [30].

Built upon the original PACIFIER, our TGBA-F extended it with three components by forecasting future events, incorporating a time-aware mechanism, and exploring a subgroups



(a) Previous work: Prediction with only OW

τ_{OW} (hours) with e_{OW} events
τ (hours) with e_{FW} events
Shock
Non-shock
Shock
Non-shock
Observation Window
(OW)
Forecasting Window
(FW)

(b) Our work: Prediction with both OW and FW

Fig. 1. Early prediction with (a) only OW and (b) both OW and FW.

$$\mathcal{R}_3(\boldsymbol{V}^i,\boldsymbol{Z}^i)=\boldsymbol{V}^i\boldsymbol{Z}^i\boldsymbol{I}$$

$$=\boldsymbol{V}^i\boldsymbol{Z}^i\begin{bmatrix}1 & & & 0\\ & 1 & & \\ & & \ddots & \\ 0 & & & 1\end{bmatrix}$$

$$=[\boldsymbol{V}_1^i-\boldsymbol{V}_2^i,\boldsymbol{V}_2^i-\boldsymbol{V}_3^i\ldots,\ \boldsymbol{V}_{e^i-1}^i-\boldsymbol{V}_{e^i}^i]$$

(a) Original Smoothness Term

$$\mathcal{R}_3(\boldsymbol{V}^i,\boldsymbol{Z}^i,\Delta\boldsymbol{T}^i)=\boldsymbol{V}^i\boldsymbol{Z}^i diag\left(iif(\Delta\boldsymbol{T}^i)\right)$$

$$=\boldsymbol{V}^i\boldsymbol{Z}^i\begin{bmatrix}iif(\Delta t_1^i) & & & 0\\ & iif(\Delta t_2^i) & & \\ & & \ddots & \\ 0 & & & iif(\Delta t_{e^i-1}^i)\end{bmatrix}$$

$$=[iif(\Delta t_1^i)(\boldsymbol{V}_1^i-\boldsymbol{V}_2^i),iif(\Delta t_2^i)(\boldsymbol{V}_2^i-\boldsymbol{V}_3^i),\ldots,iif(\Delta t_{e^i-1}^i)(\boldsymbol{V}_{e^i-1}^i-\boldsymbol{V}_{e^i}^i)]$$

(b) Time-aware Smoothness Term

Fig. 2. Comparison of (a) original smoothness term and the proposed (b) time-aware smoothness term.

basis approach. In the following sections, each component will be described in more details.

1) Forecasting Future Events:

In this work, we focus on early prediction for septic shock. To do so, we are given EHRs of a patient's visit until τ hours before an endpoint to predict whether or not this patient will develop to septic shock τ hours later. For septic shock patients, the endpoint is the onset time of septic shock; whereas for non-septic shock patients, the endpoint is the end of sequences. As shown in Figure 1 (a), from the start of a visit until τ hours before the endpoint is denoted as *observation window (OW)*, while the τ-hour leading up to the endpoint is denoted as *hold-off window (HW)*.

In previous work, only EHRs in the OW were used to predict what would happen τ hours later, as shown in Figure 1 (a); while in this work, we inferred events in a *Forecasting Window (FW)* and applied it together with the actual EHRs from OW for early prediction, as shown in Figure 1 (b). Specifically, we extracted the mapping $\hat{\mathbf{U}}$ in Eq.(1) and employed it to forecast future events for the next τ hours in FW. The forecasted events can reflect the trend of septic progression and capture more distinct latent patterns approaching the endpoint, thus they are helpful for the early prediction task.

To forecast future events in FW, we firstly need to determine the number of future events e_{FW}. To do so, we defined e_{OW} as the number of events in OW and τ_{OW} as the duration of OW in hours. Our assumption here is that the ratio of e_{FW} to the duration of τ is the same as the ratio of e_{OW} to τ_{OW}, thus we have $\frac{e_{FW}}{\tau}=\frac{e_{OW}}{\tau_{OW}}$ and then $e_{FW}=\frac{e_{OW}}{\tau_{OW}}\times\tau$. Once the number of forecasted future events was determined, their time stamps were randomly set within the τ hours. Then we applied Eq.(1) to impute all missing entries in OW and to forecast the events in FW simultaneously.

2) Time-aware Mechanism:

For a patient's visit $\mathbf{X}^i=\{x_1^i,...,x_{e^i}^i\}$, the time intervals between consecutive events in $\mathbf{X}^i$ are denoted as: $\Delta\mathbf{T}^i=[\Delta t_1^i,...,\Delta t_{e^i-1}^i]$. In EHRs, these $\Delta t_j^i, j\in[1,e^i-1]$ can vary greatly from minutes to days. To incorporate these irregular time intervals into the original Eq.(1), we rewritten the original smoothness term $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i)$ as $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i,I)$, where I is an unit matrix. Then rather than using the unit matrix I which assumes equal time intervals, we introduced a diagonal matrix $\Delta\mathbf{T}^i$. The entries of $\Delta\mathbf{T}^i$ were determined by the time intervals between two consecutive events and thus we modified the $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i)$ to be $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i,\Delta\mathbf{T}^i)$. In other words, the original $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i,I)$ controlled the variations among consecutive latent patterns equally; while in EHRs, we expected that the variation between two consecutive events with smaller time intervals should be smaller than the variation between two consecutive events with larger time intervals. As a result, time-aware mechanism could promote the robustness of data imputation to fit more general and practical situations.

More specifically, the $\Delta\mathbf{T}^i$ was converted into weights via an *i*rregular *i*nterval transformation *f*unction $iif(.)$ to constrain a more smooth longitudinal transition of latent patterns. Then the original $\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i)$ can be rewritten as:

$$\mathcal{R}_3(\mathbf{V}^i,\mathbf{Z}^i,\Delta\mathbf{T}^i)=\left\|\mathbf{V}^i\mathbf{Z}^i\mathrm{diag}(iif(\Delta\mathbf{T}^i)\right\|_F^2 \tag{3}$$

Figure 2 shows the difference between original smoothness term and our proposed time-aware smoothness term. In this work, we had $iif(\Delta\mathbf{T}^i) = exp(-\alpha\Delta\mathbf{T}^i)$, with α being a parameter to regulate the impact of $iif(.)$. By using $iff(\Delta\mathbf{T}^i)$ as diagonal, $\mathcal{R}_3(\mathbf{V}^i, \mathbf{Z}^i)$ could pose smoothness constraints to the corresponding variations between consecutive latent patterns. Specifically, when a time interval is small, we expected the consecutive latent patterns should be more similar; while when a time interval is large, the model should not punish much even if the consecutive latent patterns are very different. Consider the following two extreme cases: *a)* When an interval $\Delta t^i_j, j \in [1, e^i - 1]$ approaches 0, i.e. two consecutive events happen simultaneously, and the function $iff(\Delta t^i_j)$ tends to be 1. In this case, the smoothness regularization term poses the largest impact, and thus the missingness in latter event will be primarily inferred from its previous event; *b)* when the interval is large, the $iff(\Delta t^i_j)$ will be close to 0. In this case, impact of the smoothness regularization term is omitted and the sparseness term $\mathcal{R}_1(\hat{\mathbf{U}})$ and overfitting term $\mathcal{R}_2(\mathbf{V}^i)$ contribute more to constrain the decomposition, therefore missing entries in latter event are mainly inferred from latent patterns that generally exist among all other events.

3) Subgroup Basis Approach:

Original PACIFIER explored two approaches: the Individual Basis Approach (IBA) assumed latent patterns to be heterogeneous, therefore each patient possesses a specific mapping between features and latent patterns, i.e. $\hat{\mathbf{U}} = \mathbf{U}^i$; while the Shared Basis Approach (SBA) assumed latent patterns to be homogeneous for overall patients, i.e. $\hat{\mathbf{U}} = \mathbf{U}$.

In our proposed TGBA-F framework, the mapping between features and latent patterns are shared within subgroups, i.e. $\hat{\mathbf{U}} = \mathbf{U}^{(l)}, l \in [1, L]$, where L is the number of subgroups. Denote the size of patients' visits in subgroups as $N^{(l)}, l \in [1, L]$, the objective function Eq.(1) can be rewritten for TGBA-F as shown in Eq.(4).

$$\begin{aligned}\min_{\mathbf{S}^{(l)i},\mathbf{U}^{(l)},\mathbf{V}^{(l)i}} \quad \sum_{l=1}^{L}\Big[\sum_{i=1}^{N^{(l)}} \frac{1}{2e^{(l)i}}\mathcal{D}(\mathbf{S}^{(l)i},\mathbf{U}^{(l)},\mathbf{V}^{(l)i}) + \\ \lambda_1\mathcal{R}_1(\mathbf{U}^{(l)}) + \lambda_2\sum_{i=1}^{N^{(l)}}\frac{1}{2e^{(l)i}}\mathcal{R}_2(\mathbf{V}^{(l)i}) \qquad (4) \\ + \lambda_3\sum_{i=1}^{N^{(l)}}\frac{1}{2e^{(l)i}}\mathcal{R}_3(\mathbf{V}^{(l)i},\mathbf{Z}^{(l)i},\mathbf{T}^{(l)i})\Big] \\ \text{s.t. } \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{S}^{(l)i}) = \mathcal{P}_{\mathbf{\Omega}^i}(\mathbf{X}^{(l)i}),\ \mathbf{U}^{(l)} \geq 0\end{aligned}$$

B. Two Classifiers: LSTM and T-LSTM

As a variation of RNN, *LSTM* can retain the long-term dependencies through a gating mechanism [31]. Recently, LSTM gained popularity in biomedical domain, as 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 stamps. Each block in LSTM consists of a memory cell state and three gates: forget gate, input gate, and output gate. The three gates interact with each other to control the flow of information. More specifically, the forget gate determines what information from previous memory cell state is expired and should be removed; the input gate selects information from the candidate memory cell state to update the current 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.

Comparing to the architecture of LSTM, *T-LSTM* divides the previous memory into short-term and long-term; then it reserves the long-term memory and introduces a time-aware mechanism to adjust the short-term memory to hold. One main difference between LSTM and T-LSTM is how the memory is controlled by forget gate. In LSTM, forget gate is directly applied to the previous memory; while in T-LSTM, it acts on the previous memory with short-term memory being adjusted by the weights derived from time intervals. Both LSTM and T-LSTM were employed in this work, with the input being multivariate temporal sequence of patients, and output from the last step being used to make prediction.

IV. Experimental Setup

A. Dataset Description

This study used de-identified EHRs obtained from adult patients (age $\geqslant$ 18 years) hospitalized within the Christiana Care Health System from July 2013 to December 2015, corresponding to 119,968 unique patients and 210,289 hospitalizations. The EHRs contain both *static* and *dynamic* information. *Static* information contains patient background such as age, sex, race, etc. *Dynamic* information is collected multiple times at irregular intervals during the patients' entire hospitalization and has a time stamp associated with each record. Along with time stamps, identifiers, locations, and description, there are four categories of main attributes as follows:

- Vital signs: mean arterial pressure (MAP), systolic blood pressure (SBP), etc.
- Lab results: white blood cell count (WBC), Bands, BUN, Creatinine, Bilirubin, sedimentation rate, etc.
- Intervention: oxygen source, change of oxygen source, FiO2, drug administration, intravenous therapy, etc.
- Location: location type (emergency department, nurse, step down, intensive care unit), code.

1) Target Population & Labeling:

The study population are patients with *suspected infection* identified by the presence of any type of antibiotic, antiviral, or antifungal administration, or a positive test result of Point of Care Rapid, and it consists of 52,919 visits with 4,224,567 medical events. Note that the study population, the aforementioned rules for identifying suspected infection, and the rules for septic shock labeling in next paragraph were determined by two leading clinicians with extensive experience on this subject from Mayo Clinic and Christiana Care Health System.

Supervised models depend heavily on the accurate label of the training set. However, acquiring the true label (i.e., septic shock and non-shock) can be challenging. Although diagnosis codes, such as International Classification of Diseases, Ninth Revision (ICD-9), are widely used for clinical labeling, 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 purpose. Indeed, it has been widely argued that ICD-9 cannot be used for establishing reliable gold standards for various clinical conditions [32], [33]. More importantly, ICD-9 cannot tell when septic shock occurs at event level, which is essential for our task. On the basis of the Third International Consensus Definitions for Sepsis and Septic Shock [21], our domain experts identified septic shock as any of the following conditions are met:

- Persistent hypertension as shown through two consecutive readings ($\leq$ 30 minutes apart).
  - Systolic Blood Pressure (SBP) $<$ 90 mmHg
  - Mean Arterial Pressure (MAP) $<$ 65mmHg
  - Decrease in SBP $\geq$ 40 mmHg with an 8-hour period
- Any vasopressor administration.

When combing both ICD-9 and the domain experts' rules, we identified 1,869 shock positive visits and 23,901 negative visits. Consider the highly imbalanced ratio between positive and negative visits, we further conducted a stratified random sampling on negative visits while keeping the same underlying distribution of: age, sex, ethnicity, duration of stay and the number of events. Finally, the dataset contains 3,738 visits (1,869 positives and 1,869 negatives) with 145,421 events.

2) Time Irregularity & Missing Data Analysis:

Each patient's visit in EHRs consists of multivariate events with irregular time intervals. The time intervals between two consecutive events in our data range from 0.94 seconds to 28.19 hours. Since different features are measured at different events, plenty of missing entries exist in EHRs. For instance, vital signs are generally measured every 8 hours, while lab values are measured every 24 hours. Hence there may not be available readings for lab results when a new event is created for vital signs. Table I shows the missing rates of 14 structured features selected from the dataset in our analysis. On average, the missing rate is 80.37%.

TABLE I
RESPECTIVE MISSING RATE FOR EACH FEATURE AND THEIR AVERAGE.

Category	Feature	Missing Rate (%)
Vital Signs	Temperature	73.19
	RespiratoryRate	56.10
	HeartRate	53.41
Metabolic System	Bands	99.08
	Lactate	97.83
	WBC	93.07
	Platelet	93.07
Cardiovascular System	MAP	68.86
	SystolicBP	63.91
Respiratory System	FIO2	84.76
	PulseOx	58.26
Renal System	BUN	92.84
	Creatinine	92.84
Hepatic System	BiliRubin	97.99
Mean		**80.37**

B. Forecasting Future Events Settings: truncated vs. entire

Prior research on applying the matrix decomposition based methods for disease prediction all utilized *truncated* EHRs sequences, i.e., all events in OW shown in Figure 1 (a), to extract meaningful latent patterns $\hat{\mathbf{U}}$. However, when applying such methods to forecast future events, we have the choices of either using the *entire* training sequences or using the *truncated* training sequences to learn $\hat{\mathbf{U}}$. *Entire* means benefiting from the whole sequences of septic shock and non-septic shock visits for latent pattern extraction; while *truncated* is referring to adoption of the part of sequence included only in OW to find latent patterns. The advantage of using *entire* sequences is that the longer the sequences are, the more latent patterns can be considered and discovered; while the advantage of using *truncated* sequences is that the training data used in learning $\hat{\mathbf{U}}$ and classification stage are the same as the testing data for the classification task, thus the discovered patterns are more likely to emerge and be representative for the early diagnosis task. Thus, for learning $\hat{\mathbf{U}}$, we explored using both the ***entire (en)*** and the ***truncated (tr)*** sequences and named them as $\hat{\mathbf{U}}_{en}$ and $\hat{\mathbf{U}}_{tr}$, respectively. In both settings, we then applied the learned mapping $\hat{\mathbf{U}}_{en}$ or $\hat{\mathbf{U}}_{tr}$ to forecast future events for both training and testing data for the classification stage.

C. Subgroups

We explored different ways to partition patients into subgroups, including clustering [34] and dynamic time warping [35], etc. Much to our surprise, the best results was achieved when we using two basic demographic properties, i.e., *age* and *sex*. Moreover, some prior literature has shown that age and sex are often closely related to septic progression [36] [37]. Indeed, our preliminary analysis showed that age and sex are both highly dependent with the septic shock. Specifically, we discretetized age into 4 subintervals with the breakpoints of $\{30, 50, 70\}$ and χ^2 tests were performed to examine the relationship between age and septic shock. We found that there is a significant difference on the ratio of sepsis shock among the four age groups: $\chi^2(3, N = 3,738) = 307.38, p \approx 0 \ll 0.01$, in that older patients are more likely to develop septic shock than younger ones. Similarly, there is a significant difference on the ratio of septic shock between Female and Male: $\chi^2(1, N = 3,738) = 7.58, p = 0.0059 < 0.01$ in that Male is more likely to develop septic shock. Combining the age with sex, we got a total of 8 subgroups.

D. Setups

We conducted a series of experiments to evaluate the effectiveness of proposed TGBA-F for septic shock early prediction task using two classifiers: LSTM first and then T-LSTM.

For LSTM, we first explored the effectiveness of the original IBA and SBA with forecasted future events induced by either $\hat{\mathbf{U}}_{tr}$ or $\hat{\mathbf{U}}_{en}$, and compared them against five baseline methods without forecasted events: *CF* and *MEAN*, *SVDImpute*, and the original *IBA* and *SBA*.

- Carrying forward (CF): fills the missing values with the last observation until the next value is observed;
- Mean imputation (MEAN): fills all missing values with the mean value of the corresponding feature;
- SVDImpute: independently treats all events from different visits, stacks them into one matrix and then runs SVD [9];
- IBA: is Individual Basis PACIFIER Approach that learns latent patterns for each *individual* patient [11];
- SBA: is Shared Basis PACIFIER Approach that learns latent patterns for the *entire* population [11].

Note that $\hat{\mathbf{U}}_{en}$ is not applicable for IBA because IBA is implemented individually for each patient, thus we can only learn $\hat{\mathbf{U}}_{tr}$ for IBA. Our results showed that using *Forecasted* future events could improve the effectiveness of IBA and SBA in that IBA-F$_{tr}$ and SBA-F$_{en}$ achieved better overall performance. As a result, we employed *truncated* sequences for IBA and *entire* sequences for SBA hereinafter. Next, with time-aware mechanism, our results showed that Time-aware IBA-F (TIBA-F) using *truncated* sequences and Time-aware SBA-F (TSBA-F) using *entire* sequences were indeed more effective than IBA-F$_{tr}$ and SBA-F$_{en}$ without time-aware mechanism. Finally, we explored the effectiveness of subgroup basis approach with different ways of grouping patients according to their *age (a)* and *sex (s)*. The results showed that the best performance was achieved by our proposed TGBA-F, which combines all three components: forecasting future events with $\hat{\mathbf{U}}_{en}$ learned from *entire* training sequences, incorporating time-aware mechanism, and taking subgroup basis approach based on both *age* and *sex* (*a*&*s*).

To further evaluate the effectiveness of time-aware mechanism, we compared GBA-F (TGBA-F *without* time-aware) and TGBA-F with LSTM vs. GBA-F and TGBA-F with T-LSTM. Herein, our motivation is to investigate whether the latent patterns extracted by TGBA-F can be used to further improve the performance of T-LSTM. In this experiment, we forecasted future events using *entire* training sequences and took subgroup basis approach based on (*a*&*s*).

To determine the optimal number of latent patterns in all matrix decomposition based methods (i.e., TGBA-F, TIBA-F, TSBA-F, etc.), we ran grid search and determined the optimal number as 8 for all methods. For time-aware mechanism, we used grid search to investigate the optimum value for the parameter α in the function $iif(\Delta\mathbf{T}^i) = exp(-\alpha\Delta\mathbf{T}^i)$ and our results showed that the optimum value was 0.2, which enabled the $iif(\Delta\mathbf{T}^i)$ to have a nearly full distribution over the range of [0,1]. For both LSTM and T-LSTM, we used one hidden layer with 50 hidden neurons and 514 maximum sequence length. We adopted the Adam optimizer [38] with the batch size 50 and 30 epochs, and early stopping was employed with 5 patience after minimum 10 epochs.

E. Evaluation Metrics

Metrics of accuracy (Acc), recall, precision (Prec), F-score (F1) and area under the ROC curve (AUC) were employed for evaluating our models. Accuracy is the proportion of patients whose labels are correctly identified. Recall indicates what proportion of patients that actually have septic shock can be correctly diagnosed by the model as septic shock. Precision tells what proportion of patients who are diagnosed as septic shock actually have septic shock. F1 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 and AUC to compare different models. All models were evaluated by 5-fold cross validation.

V. Results

A. Effectiveness of Three Components in Early Prediction

1) Performances of $\tau = 4$ and Average of $\tau \in [1, 8]$:

Based on the three components in TGBA-F, we divided experiments into Baseline and three sub-sessions: *1) F*orecasting future events (**F**); *2)* appending the *T*ime-aware mechanism to (F), i.e., (**FT**); and *3)* appending the sub*G*roup basis approach to (FT), i.e., (**FTG**). Table II shows the performance on two prediction tasks: prediction *Task I* (4-hour-before shock prediction) and *Task II* (1-to-8-hour-before shock prediction when the hold-off window varying from 1 hour to 8 hours by 1 hour increment). Specifically, the first column indicates the sub-session; the second column is the missing data handling method; columns 3 to 7 present our evaluation metrics for the prediction *Task I*; and columns 8 to 12 present average (mean ± std) performance for the prediction *Task II*. For each sub-session, the best results are marked in bold; in sub-session (**F**), the best results are underlined; in sub-session (**FT**), the best results are labeled with $*$; and the best performance in sub-session (**FTG**) are labeled with both underline and $*$. Finally, the best model across all sub-sessions are shaded. In the following, we will report the performance in each sub-session and then compare across them. Table II (Baseline) shows that both IBA and SBA outperformed other three baseline methods except that SVDImpute had the best precision among them. IBA had slightly better performance than SBA.

(F) *with Forecasted Future Events*:

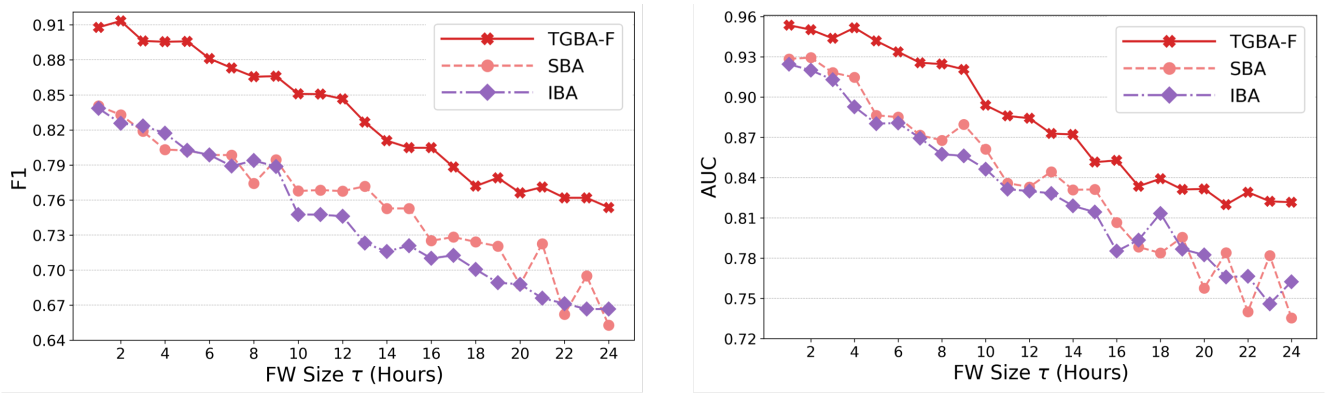
We explored whether the performance of IBA and SBA could be improved with *Forecasted* future events using *truncated* sequences, denoted as IBA-F$_{tr}$ and SBA-F$_{tr}$ respectively, and using *entire* sequence for SBA only, denoted as SBA-F$_{en}$. As discussed previously, *entire* sequence is not applicable for IBA. Table II (**F**) shows that all three forecasted models outperformed the five baseline methods for both *Task I*

TABLE II

COMPARISON OF DATA IMPUTATION METHODS FOR BASELINE METHODS AND THREE SUB-SESSIONS: (F) WITH FORECASTED FUTURE EVENTS; (FT) WITH TIME-AWARE MECHANISM; AND (FTG) WITH SUBGROUP BASIS APPROACH.

Sub-session	**Method**		***Task I*: $\tau = 4$**					***Task II*: Overall (mean ± std) of $\tau \in [1, 8]$**				
			Acc	Recall	Prec	F1	AUC	Acc	Recall	Prec	F1	AUC
Baseline	CF		.824	.731	.842	.783	.887	.813±.029	.775±.031	.802±.040	.787±.025	.888±.020
	MEAN		.804	.800	.761	.780	.892	.801±.028	.747±.055	.805±.028	.774±.033	.881±.024
	SVDImpute		.804	.674	**.843**	.749	.884	.783±.027	.741±.043	.769±.045	.753±.020	.865±.022
	IBA		**.831**	**.869**	.772	**.817**	.893	**.831±.023**	.811±.029	**.814±.041**	**.811±.016**	.892±.023
	SBA		.816	.863	.751	.803	**.915**	.827±.024	**.818±.029**	.802±.034	.809±.020	**.900±.024**
(F)	IBA-F_{tr}		**.850**	.869	.831	.849	.911	**.848±.012**	.842±.036	**.852±.028**	**.846±.013**	**.909±.019**
	SBA-F_{tr}		.844	.840	.840	.840	.919	.834±.021	.831±.029	.835±.026	.833±.021	.905±.021
	SBA-F_{en}		**.850**	**.874**	.827	**.850**	**.922**	.845±.024	**.846±.015**	.845±.037	.845±.021	.909±.022
(FT)	TIBA-F		**.878***	**.903***	.854	**.878***	.912	**.874±.015***	.864±.020	**.881±.025***	**.872±.014***	.917±.018
	TSBA-F		.872	.846	**.886***	.866	**.928***	.867±.021	**.872±.032***	.863±.035	.867±.020	**.921±.018***
(FTG)	TGBA-F	(*a*)	.892	**.914***	.870	.891	.939	.881±.022	**.891±.026***	.873±.032	.881±.022	.928±.020
		(*s*)	.883	.863	.894	.878	.937	.878±.022	.862±.034	.892±.039	.876±.021	.930±.013
		(*a&s*)	**.903***	.857	**.938***	**.896***	**.952***	**.894±.016***	.874±.023	**.910±.025***	**.891±.015***	**.941±.011***

- The bold numbers in **Baseline** are the best results among baseline methods;
- The bold numbers with underline in sub-session (**F**) are the best results with *truncate (tr)* or *entire (en)* sequences;
- The bold numbers with * in (**FT**) are the best results with time-aware mechanism. TIBA-F uses *truncated* sequences and TSBA-F uses *entire* sequences;
- The bold numbers with underline and * in (**FTG**) are the best results with *entire* sequences, time-aware mechanism, and subgroup basis approach based on: *age* (*a*), *sex* (*s*) and *age* & *sex* (*a&s*);
- The shaded cells have the best performance among all sub-sessions.



(a) Comparison of F1 for TGBA-F vs. SBA & IBA

(b) Comparison of AUC for TGBA-F vs. SBA & IBA

Fig. 3. Comparison of TGBA-F (using *entire* sequences and grouping patients by (*a&s*)) and baseline methods (SBA & IBA) over a FW of 24 hours.

(except precision) and *Task II*. Among the three models, SBA-F_{en} has the best F1 and AUC performance for the prediction *Task I*, and IBA-F_{tr} have the best F1 and AUC performance for the prediction *Task II*. Note that SBA-F_{en} outperformed SBA-F_{tr} for both prediction tasks, thus we only considered *entire* sequences for SBA in the following experiments.

*(**FT**) with Time-aware Mechanism*:

We evaluated the effectiveness of the time-aware mechanism on the best models so far, i.e., IBA-F_{tr} and SBA-F_{en}, denoted as TIBA-F and TSBA-F respectively. The results in Table II (**FT**) shows that both TIBA-F and TSBA-F outperformed IBA-F_{tr} and SBA-F_{en} on *Task I* and *Task II*. Additionally, there is no clear winner between TIBA-F and TSBA-F in that TIBA-F has better F1, while TSBA-F has better AUC for both prediction tasks.

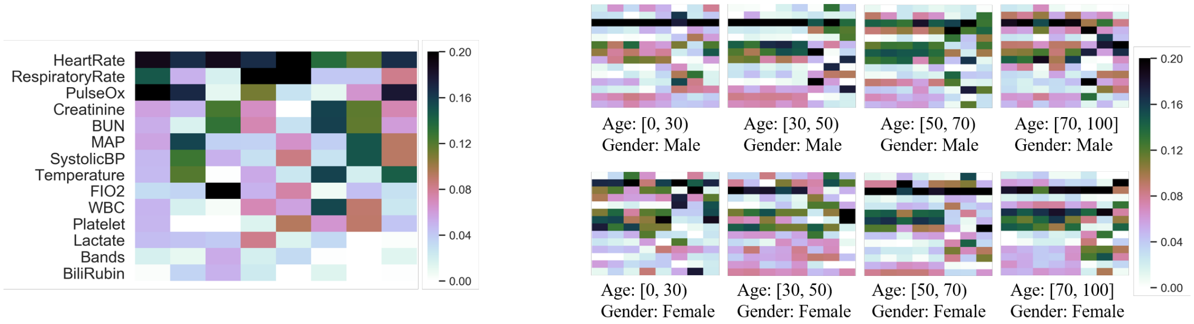
*(**FTG**) with Subgroup Basis Approach*:

To evaluate our subgroup basis approach, we compared it against the best IBA model, i.e. TIBA-F, and the best SBA model, i.e. TSBA-F, so far. To do so, we used *entire* sequence setting and explored subgroup basis approach using three different ways of partitioning the patients: (*a*) 4 groups by age; (*s*) 2 groups by sex; and (*a&s*) 8 groups by age and sex

TABLE III
GBA-F vs. TGBA-F & LSTM vs. T-LSTM using *entire* training sequences and grouping patients by ($a\&s$).

Method	**Classifier**	***Task I***: $\tau = 4$					***Task II***: Overall (mean ± std) of $\tau \in [1, 8]$				
		Acc	Recall	Prec	F1	AUC	Acc	Recall	Prec	F1	AUC
GBA-F	LSTM	.872	.863	.873	.868	.930	.869±.019	.865±.030	.872±.037	.868±.018	.928±.015
TGBA-F	LSTM	.903	.857	**.938**	.896	.952	.894±.016	.874±.023	.910±.025	.891±.015	.941±.011
GBA-F	T-LSTM	.889	.903	.873	.888	.949	.886±.019	.881±.021	.889±.030	.884±.018	.939±.017
TGBA-F	T-LSTM	**.911**	**.914**	.904	**.909**	**.961**	**.907±.011**	**.901±.022**	**.912±.016**	**.906±.012**	**.953±.012**

- The bold numbers have the best performance.



(a) Overall shared mapping $\widehat{U}_{en}$ decomposed from TSBA-F

(b) Sub-group shared shared mapping $\widehat{U}_{en}$ decomposed from TGBA-F

Fig. 4. Visualization of mapping $\hat{\mathbf{U}}_{en}$ achieved from: (a) *TSBA-F* for overall patients; and (b) *TGBA-F* for subgroups. Notably, the columns in different $\hat{\mathbf{U}}_{en}$ indicate different latent patterns.

collectively. Table II (**FTG**) shows all three subgroup basis approaches outperformed TSBA-F and TIBA-F in terms of F1 and AUC on both prediction tasks. Additionally, grouping patients by either age or sex achieved similar results. When combining age and sex together, the performance could be further improved. Our results suggest that different age and sex subgroups possess different latent patterns, therefore group basis approach can not only achieve more specific latent patterns for each group, but also provide more exact clues for data imputation from patients with similar properties.

2) *Performance of $\tau \in [1, 24]$:*

In order to fully evaluate the performance of TGBA-F which achieved the best performance so far, using *entire* sequences for forecasting, incorporating the time-aware mechanism, and grouping patients by ($a\&s$), it is further compared with the two original PACIFIER approaches without any extensions: i.e., IBA and SBA. Figure 3 shows that TGBA-F consistently outperformed the original IBA and SBA on a 1-to-24-hour-before shock prediction task. When τ becomes larger, it is more challenging to early predict the septic shock. In this case, the two baseline methods fluctuated heavily, while our TGBA-F stays much more stable.

B. GBA-F vs. TGBA-F & LSTM vs. T-LSTM

To further evaluate the effectiveness of time-aware mechanism, we compared GBA-F (TGBA-F *without* time-aware) and TGBA-F with LSTM vs. GBA-F and TGBA-F with T-LSTM. Both GBA-F and TGBA-F use *entire* sequences and group patients by ($a\&s$). Table III shows the predictive performance on *Task I* (4-hour-before) & *Task II* (1-to-8-hour-before shock prediction). Overall, for both GBA-F and TGBA-F, T-LSTM outperformed LSTM and thus T-LSTM is indeed more suitable for modeling EHRs. On the other hand, for both LSTM and T-LSTM, TGBA-F outperformed GBA-F, thus our time-aware mechanism is indeed more effective for imputing missing data. As a result, we assumed that our time-aware mechanism and T-LSTM could tackle irregular time intervals from two different perspectives so that the best performance was generated by combining the two approaches: TGBA-F and T-LSTM.

C. Visualization of Decomposed Latent Patterns

In Figure 4, we visualized the mapping $\hat{\mathbf{U}}_{en}$ that extracted from TSBA-F for overall patients (Figure 4 (a)) and TGBA-F for each subgroup (Figure 4 (b)). Herein, the rows indicate features and columns are latent patterns. Features are sorted by their summing of weights to all latent patterns, and ordered from the largest (top) to the smallest (bottom). The darker the color, the more contribution for the feature in representing the corresponding latent pattern. Figure 4 shows overall shared mapping is different from subgroups, and also different subgroups have heterogeneous mappings.

VI. Conclusion

In this paper, based on matrix decomposition, we proposed a TGBA-F method which consists of three components including: forecasting future events, time-aware mechanism, and subgroup basis approach. The imputed data with forecasted future events can be fed into deep learning classifiers, LSTM and T-LSTM, for septic shock early prediction. Through experiments, we demonstrated that our proposed TGBA-F approach can significantly improve the performance of early prediction comparing to the baseline methods. In future works, we will explore other applicable approaches to partition subgroups. We will also analyze the implications for the decomposed latent patterns to find out their clinical meanings. Besides, a larger set of features will be introduced in future analysis to explore more sophisticated latent patterns.

VII. Acknowledgement

This paper was supported by the following NSF Grants: 1) S.E.P.S.I.S.: Sepsis Early Prediction Support Implementation System (1522107); 2) MetaDash: A Teacher Dashboard Informed by Real-Time Multichannel Self-Regulated Learning Data (1660878); and 3) CAREER: Improving Adaptive Decision Making in Interactive Learning Environments (1651909).

References

[1] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annual review of public health*, vol. 36, pp. 345–359, 2015.

[2] F. Cismondi, A. Fialho, S. Vieira, S. Reti, J. Sousa, and S. Finkelstein, "Missing data in medical databases: Impute, delete or classify," *AI in Medicine*, vol. 58, May 2013.

[3] J. Galimard, S. Chevret, C. Protopopescu, and M. RescheRigon, "A multiple imputation approach for mnar mechanisms compatible with heckman's model," *Statistics in medicine*, vol. 35, February 2016.

[4] D. Rubin and R. Little, *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons, 1987.

[5] Z. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series," *JMLR*, vol. 56, 2016.

[6] Y.-J. Kim and M. Chi, "Temporal belief memory: Imputing missing data during rnn training.," in *IJCAI*, pp. 2326–2332, 2018.

[7] P. Garca-Laencina, P. Abreu, M. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Computers in Biology and Medicine*, vol. 59, April 2015.

[8] B. K. Beaulieu-Jones and J. H. Moore, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Pacific Symposium on Biocomputing 2017*, pp. 207–218, World Scientific, 2017.

[9] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[10] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of machine learning research*, vol. 11, no. Aug, pp. 2287–2322, 2010.

[11] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144, ACM, 2014.

[12] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of parkinsons disease: variation by age, gender, and race/ethnicity," *American journal of epidemiology*, vol. 157, no. 11, pp. 1015–1022, 2003.

[13] J. Cohen, J.-L. Vincent, N. K. Adhikari, F. R. Machado, D. C. Angus, T. Calandra, K. Jaton, S. Giulieri, J. Delaloye, S. Opal, *et al.*, "Sepsis: a roadmap for future research," *The Lancet infectious diseases*, vol. 15, no. 5, pp. 581–614, 2015.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, December 1997.

[15] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder- decoder for statistical machine translation," in *EMNLP*, p. 17241734, 2014.

[16] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74, ACM, 2017.

[17] Y. Bengio and F. Gingras, "Recurrent neural networks for missing or asynchronous data," in *NIPS*, pp. 395–401, 1995.

[18] V. Tresp and T. Briegel, "A solution for missing data in recurrent neural networks with an application to blood glucose prediction," in *NIPS*, pp. 971–977, 1997.

[19] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *CoRR*, vol. abs/1606.01865, 2016.

[20] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased lstm: Accelerating recurrent network training for long or event-based sequences," in *Advances in Neural Information Processing Systems*, pp. 3882–3890, 2016.

[21] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[22] G. S. Martin, D. M. Mannino, S. Eaton, and M. Moss, "The epidemiology of sepsis in the united states from 1979 through 2000," *New England Journal of Medicine*, vol. 348, no. 16, pp. 1546–1554, 2003.

[23] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[24] G. Polat, R. Ugan, E. Cadirci, and Z. Halici, "Sepsis and septic shock: Current treatment strategies and new approaches," *EJM*, vol. 49, 2017.

[25] B. K. Beaulieu-Jones, D. R. Lavage, J. W. Snyder, J. H. Moore, S. A. Pendergrass, and C. R. Bauer, "Characterizing and managing missing structured data in electronic health records: Data analysis," *JMIR medical informatics*, vol. 6, no. 1, 2018.

[26] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

[27] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 198–206, SIAM, 2017.

[28] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 30–41, Springer, 2016.

[29] H.-G. Kim, G.-J. Jang, H.-J. Choi, M. Kim, Y.-W. Kim, and J. Choi, "Recurrent neural networks with missing information imputation for medical examination data prediction," in *BigComp*, pp. 317–323, February 2017.

[30] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *ASU*, vol. 21, 2011.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] J. Ho, C. Lee, and J. Ghosh, "Septic shock prediction for patients with missing data," *Management Information Systems*, vol. 5, April 2014.

[33] Y. Zhang, C. Lin, M. Chi, J. Ivy, M. Capan, and J. M. Huddleston, "Lstm for septic shock: Adding unreliable labels to reliable predictions," in *IEEE International Conference on Big Data*, pp. 1233–1242, 2017.

[34] M. J. Sewitch, K. Leffondre, and P. L. Dobkin, "Clustering patients according to health perceptions: relationships to psychosocial characteristics and medication nonadherence," *Journal of psychosomatic research*, vol. 56, no. 3, pp. 323–332, 2004.

[35] M. Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, pp. 69–84, 2007.

[36] J. Schröder, V. Kahlke, K.-H. Staubach, P. Zabel, and F. Stüber, "Gender differences in human sepsis," *Archives of Surgery*, vol. 133, no. 11, pp. 1200–1205, 1998.

[37] N. Nasir, B. Jamil, S. Siddiqui, N. Talat, F. A. Khan, and R. Hussain, "Mortality in sepsis and its relationship with gender," *Pakistan journal of medical sciences*, vol. 31, no. 5, p. 1201, 2015.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.