MuLan: Multilevel Language-based Representation Learning for Disease Progression Modeling

Hyunwoo Sohn*, Kyungjin Park† and Min Chi‡
Department of Computer Science, *North Carolina State University*
Raleigh, NC, USA
{hsohn3*, kpark8†, mchi‡}@ncsu.edu

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377829

***Abstract*—Modeling patient disease progression using Electronic Health Records (EHRs) is crucial to assist clinical decision making. In recent years, deep learning models such as Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) have shown great success in handling sequential multivariate data, such as EHRs. Despite their great success, it is often difficult to *interpret and visualize patient disease progression* learned from these models in a meaningful yet unified way. In this work, we present *MuLan*: a Multilevel Language-based representation learning framework that can automatically learn a *hierarchical representation* for EHRs at *entry*, *event*, and *visit* levels. We validate *MuLan* on modeling the progression of an extremely challenging disease, septic shock, by using real-world EHRs. Our results showed that these *unified* multilevel representations can be utilized not only for interpreting and visualizing the latent mechanism of patients' septic shock progressions but also for early detection of septic shock.**

***Index Terms*—Electronic health records, disease progression modeling, interpretability, representation learning**

I. Introduction

Disease Progression Modeling (DPM) [1] is a task of monitoring disease developing process and predicting future risks based on patients' historical information. DPM is crucial for understanding disease prognosis, assisting early diagnosis, and developing timely personalized interventions [2]–[9]. In recent years, the broad adoption of Electronic Health Records (EHRs) in medical systems has promoted the development of various computational methods for DPM-related tasks [5], [7], [10]–[13]. Among them, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) have become the state-of-the-art approaches. More specifically, Recurrent Neural Network (RNN) and its variants such as Long Short-Term Memory (LSTM) have demonstrated their ability to capture the *temporal dependency* over a patient's trajectory [5], [7], [11]–[13], while Convolutional Neural Network (CNN) have shown great success in capturing time-invariant *local patterns* in a short-term period of time [14]–[18]. Due to their distinct but complementary power, RNN and CNN often create a synergistic effect when combined. However, a significant barrier remains when applying them for modeling EHRs as it is often difficult to *interpret and visualize patient disease progression* in a meaningful yet unified way.

Conversely, by extracting salient features from raw data, *representation learning* can generate highly *interpretable yet effective* models across a wide range of tasks when combined with deep learning methods. As a pioneer, distributed representation learning of words has received great attention for fully capturing semantics and syntax of natural language and encoding them into a vector space [19], [20]. Furthermore, representations for various concepts (e.g., users and items in e-commerce) have been learned using RNN and CNN [21]–[23]. It is shown that representation learning can not only induce accurate predictive models but interpret the underlying relationships within the data. In healthcare, motivated by the fact that physicians can efficiently interpret patients' condition from condensed natural language-based clinical notes and patients' medical records, there has been a growing interest in adopting representation learning, especially natural language processing (NLP) methods for analyzing clinical notes and medical concepts (e.g., diagnosis codes) [4], [24], [25]. Until recently, various representation learning approaches have been successfully applied to EHRs to facilitate varied types of interpretation, for example, providing influential features or knowledge of comorbidity.

EHRs are a large-scale, systematic, and comprehensive collection of temporal health information of patients. Despite the richness of information, it is often very challenging to work with raw EHRs due to their sparse, irregular, and asynchronous nature. More specifically, EHRs can be seen as a *multilevel* structure: at the top level, EHRs are composed of *visits*; each patient's ***visit*** is composed of a sequence of *events* which reflect the health status change of the patient over time during the visit and events are commonly acquired with *irregular intervals*; each ***event*** is composed of a multivariate vector where the ***entries*** are patients' health records on corresponding features such as vital signs or lab results. In other words, a patient's visit comprises a set of events recorded at irregular times while an event contains a set of medical data entries collected at the same time. As a result, *heterogeneous relations* among data *entries*, *temporal dependencies* among *events*, and finally the *hierarchical structure* of a *visit* shape a unique characteristic of EHRs. While prior representation learning work on EHRs has mainly focused on visit-level, as far as we know, no previous work has explored a *unified* representation learning using raw EHRs across three levels.

In this work, we propose *MuLan*, a **Mu**ltilevel **Lan**guage-based representation learning framework, which can systemically learn a *hierarchical representation* for EHRs at *entry*, *event*, and *visit* levels by capturing informative temporal

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

progression patterns of a target disease in a patient's trajectory. *MuLan* can automatically summarize sparse and noisy EHRs and learn the underlying mechanism of disease progression from the summary. More specifically, *MuLan* captures three types of relational knowledge that are inherent to the hierarchy within EHRs by employing CNN and LSTM: 1) intra-event relationship among medical data entries (e.g., readings of vital signs), 2) intra-event co-occurrence patterns, and 3) inter-event temporal dependencies.

We validate *MuLan* on the task of early prediction of septic shock. Sepsis, a systemic inflammatory response to infection, is a life-threatening organ dysfunction [26]. It is a leading cause of death in U.S. hospitals where it has an impact on approximately 1.7 million adults in the U.S. each year and causes more than 250,000 deaths [27]. ***Septic shock***, the most severe complication of sepsis, is a significant health problem that can lead to a mortality rate as high as 50% and an increasing annualized incidence [28]. Moreover, each hour of treatment delay increases the mortality rate by 7.6% [29]. Prior studies have demonstrated that early diagnosis and rapid treatment in the early stage of sepsis progression are critical since they can prevent 80% of deaths [29]–[32]. However, the early diagnosis still remains challenging because no specific indicative biomarkers have yet been discovered due to its subtle but rapid progression in the early stage of the infection [33]. Specifically, the hallmark of sepsis – infection – can also be observed as the first stage of many other conditions.

Our experimental results on real-world EHRs show that by simultaneously conducting representation learning and prediction tasks, the extracted multilevel representations can be utilized not only for early detection of septic shock but also for interpreting and visualizing the latent mechanism of patients' sepsis progressions.

To summarize, this work makes the following contributions:

- To the best of our knowledge, this is the first work to propose a language-based representation learning framework that learns a hierarchical representation of raw EHRs at *entry*, *event*, and *visit* levels.
- Our framework outperforms the baselines on an extremely challenging task, septic shock early prediction.
- Our multilevel representations provide interpretation and visualization of a patient's disease progression in a meaningful yet unified way.

II. Related Work

Representation Learning in NLP: Representation learning has been extensively studied in the NLP field. Classic methods consider each word as an independent discrete concept, whereas neural network-based representation learning methods connect words by considering semantic and syntactic relations *embedded* among them in the form of a continuous vector, and thus, they are often referred to as *word embeddings*. *Word2Vec* [19] is one of the most widely used techniques, built upon the *Distributional hypothesis* [34], where words that occurred in the same contexts tend to have similar meanings. *Word2Vec* provides two representation learning algorithms: 1) Skip-gram, which tries to maximize the probability of observing context words given a center word, and 2) Continuous Bag of Word (CBOW), which maximizes the probability of observing a center word given its context words. The resulted embedding's ability to capture subtle relationships between words has been demonstrated using word analogy tasks [19]. Therefore, *Word2Vec* has been widely applied as a basis of many concept representation learning methods in several domains including healthcare [4], [25], [35] and e-commerce [36], [37]. Furthermore, Kim et al. [38] proposed a CNN-based method that can fine-tune these pre-trained embeddings. In their work, an input is initialized with pre-trained embeddings and fine-tuned in the process of a supervised task (e.g., opinion polarity detection), by absorbing the most relevant information specific to the task. Their work demonstrates the effectiveness of fine-tuning in both interpretation and classification.

Representation Learning of EHRs: In the healthcare domain, researchers have also explored the effectiveness of word representation models on a discrete type of medical data. Nguyen et al. [25] proposed *Deepr* that considers medical records as a sequence of words (i.e., medical codes) and detects predictive clinical motifs using a 1D convolutional network and logistic regression. The detected motifs are utilized to understand comorbidity, care patterns, and disease progression. Note that in their work, event-level temporal dependencies in EHRs are not considered. In our prior work, Lin et al. [39] introduced a facial image-based representation learning method, which employs CNN and LSTM to encode EHRs' temporal dependencies and local patterns into a sequence of facial images with evolving emotional expressions. The learned representations can provide the visualization of disease progression and enhance the performance of septic shock early prediction. However, the mappings from the sepsis stages to emotion expression are rather arbitrary and domain experts find it difficult to interpret and understand the resulted images. More recently, Transformer [40]-based word representation models such as BERT [41] have been widely applied to EHRs because of its training efficiency and ability to learn contextualized embeddings. Much of such work [42]–[44] utilizes visit-level discrete medical codes or clinical notes except for the works that incorporate inpatient data [45], [46]. Note that the transformer-based works mainly focus on effective predictive models while in this work, building an effective predictive model is only one of the aspects and a more important goal is to interpret and visualize the multivariate time-series EHRs through learning a unified hierarchical representation.

Closely related to this work, Choi et al. [4] proposed *Med2Vec* with a basis of Skip-gram to learn code-level and visit-level representations simultaneously from medical concepts and showed great success. *Med2Vec* incorporates information derived from the hierarchical structure of EHRs: sequential order of visits and co-occurrence of the medical codes within each visit. The learned representations not only bring improvement in prediction tasks but also provide clinically meaningful interpretations such as discovering comorbidity. In

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

this work, *Med2Vec* is one of our baseline models. Our work differs from *Med2Vec* in the following aspects: 1) granularity of EHRs: *MuLan* targets fine-grained EHRs (i.e., event-level) which is noisy, sparse, and heterogeneous while *Med2Vec* uses coarse-grained EHRs (i.e., visit-level); 2) type of learning: *MuLan* learns more specific representation to a target task by incorporating supervised task compared to *Med2Vec*, which employs an unsupervised approach solely.

Disease Progression Modeling using EHRs: EHRs have been a popular research platform with increasing availability to develop predictive models for disease progression [47]–[49], phenotyping [2], [50], [51], and diagnosis prediction [7], [8], [52]. However, EHRs also pose numerous challenges due to their noisy, fragmental, and high dimensional nature. To this end, the incorporation of deep learning techniques can assist in mitigating the impact of noise and learning complex relationships among medical events. For example, RNN is widely used for modeling multivariate time series data with missing values [53]. LSTM, one of the RNN-variants, has been implemented for general diagnosis such as phenotyping [12] or disease progression modeling [47]. Despite their great success, it is often hard to *interpret and visualize patient disease progression* in a meaningful yet unified way. *Interpretability and visualization* of computational models are extremely critical in healthcare-related domains. In real hospital settings, it is generally more important to learn discriminative *interpretable* patterns that capture an informative progression of a disease than to induce an accurate predictive computational model, and equally importantly, visualizing patient disease progression is the key for timely medical interventions and treatment. Therefore, by combining multilevel representation learning together with deep learning, we expect *MuLan* would not only lead to an effective predictive model but also result in an interpretable and visualizable clinically reasonable model.

III. Proposed Framework

Our dataset can be represented as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, where N is the total number of hospital visits. It is composed of multivariate irregular time series data and each visit $\boldsymbol{x}_k$ consists of a sequence of events: $\boldsymbol{x}_k = \{\boldsymbol{x}_k^1, ..., \boldsymbol{x}_k^{T_k}\}$, where $\boldsymbol{x}_k^t$ represents patient's records at timestamp t and T_k is the number of events in $\boldsymbol{x}_k$, which varies across different visits. We have $\boldsymbol{x}_k^t \in \mathbb{R}^D$, where D is the number of medical data entries recorded at each event. For each $\boldsymbol{x}_k$, we are provided with an output label $y_k = \{1, 0\}$ which represents septic shock and non-septic shock, respectively. Our goal in this work is to learn a function for the septic shock early prediction combining with a fine-training process for the pre-trained representation. Note that the visit index k is omitted in the following sections for easier illustration.

Our proposed framework consists of three stages: 1) ***EHRs Summarization***, which generates sets of temporal summaries from EHRs, 2) ***Entry-level Representation Learning***, which learns distributed language representations from the generated summaries, and 3) ***Event and Visit-level Representation Learning***, which fine-tunes the entry-level representation and learns event and visit-level representations while making septic shock early prediction. The three stages will be described individually in the following sections.

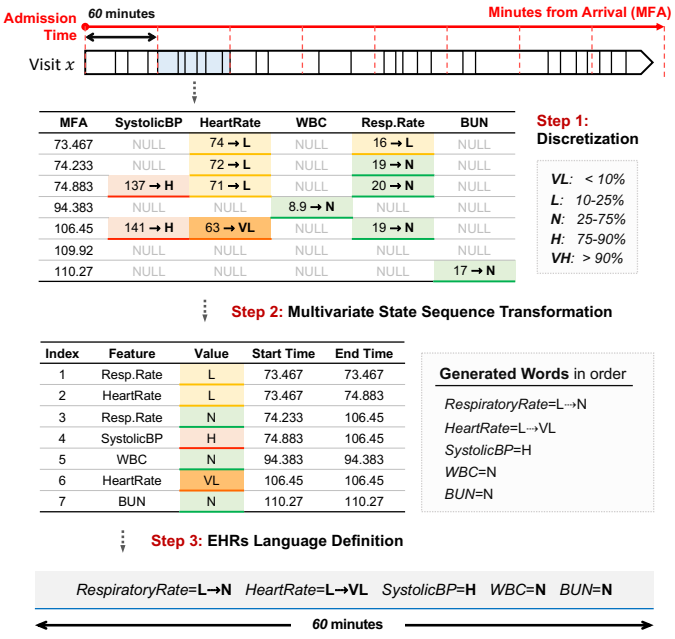

Fig. 1. EHRs Summarization Stage

A. EHRs Summarization

Fig. 1 shows detailed steps of the EHRs summarization stage. A hollow arrow on the top of Fig. 1 indicates a patient's visit $\boldsymbol{x}$ composed of a sequence of events, which is represented as a set of lines within the arrow. The objective of this stage is to aggregate a set of events in a *specific time window* into language-based representation. To do so, visit $\boldsymbol{x}$ with T events is divided into a series of L fixed time intervals where $L < T$, and the events in each time interval are summarized with the following three steps: 1) Discretization, 2) Multivariate State Sequence Transformation, and 3) EHRs Language Definition, as shown in Fig. 1.

Discretization: The continuous entries are first transformed into discrete symbols using a subgroup-based method described in [54]. Patient subgrouping is introduced to alleviate the distinct reference ranges of physiological parameters for different patient groups (e.g., The normal range of heart rate for young adults is different from the one for aged adults). For each patient group, 10%, 25%, 75%, and 90% percentiles are used to discretize the numeric values as defined in [55]. From the lowest percentile range 0-10%, we map 5 categories of *very low* (VL), *low* (L), *normal* (N), *high* (H), and *very high* (VH) to indicate the state of each feature.

Multivariate State Sequence Transformation: Following Batal et al.'s approach [55], we transform the discretized records from Step 1 into Multivariate State Sequences (MSSs). State S is defined as (F, V) where F is a temporal feature and V is the discretized value for the corresponding feature. A state interval I is denoted as (F, V, s, e) where s and e

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

denote start and end time of the state S. For example, the fourth state interval (*SystolicBP*, H, 74.883, 106.45) in Fig. 1 indicates that the patient's systolic blood pressure maintained high from minutes 74.883 to 106.45. After every state interval I is recorded, we sort them by their start and end time to keep temporal relations inside, and denote them as Z^i, the MSS at i-th time interval. In sum, the events in a i-th time interval are transformed into a set of state intervals $\{{I_i}^1, ..., {I_i}^{m^i}\}$ where m^i is the number of state intervals, and form Z^i. As a result, a visit $\boldsymbol{x}$ with L time intervals can be represented as **a set of *Multivariate State Sequences*** (MSSs):

$$\boldsymbol{x} \approx \boldsymbol{MSSs} = \{Z^1, \cdots, Z^L\}$$

EHRs Language Definition: In this step, we generate an EHR-based language by defining a word for each medical entry. As shown in Fig. 1, a word is formed as a combination of a feature F and its value V from each state interval ${I_i}^k$, $k \in [1, m^i]$, which can represent either a steady-state or a trend depending on the changes that occur within the i-th time interval. For example, "*SystolicBP=H*" indicates that systolic blood pressure maintained as *high* while "*Temperature=L→H*" indicates a patient's body temperature elevated from *low* to *high*. Once we transform every state interval in the MSS into words, we can generate a complete form of a summary by piecing the words together while keeping their temporal order. S^i is the summary of the i-th MSS Z^i or time interval, and the number of words n^i in S^i varies according to the number of features recorded in i-th time block. Finally, a visit $\boldsymbol{x}$ can be represented as ***a summarized document(Doc)***:

$$\boldsymbol{x} \approx \boldsymbol{Doc} = \{S^1, \cdots, S^L\}, \text{ where } S^i = \text{“}w_1 + \cdots + w_{n^i}\text{”}$$

B. Entry-level Representation Learning

CBOW [19] is adopted in this stage to capture heterogeneous relations among our medical words based on their co-occurrence information. The major difference between our medical summary and a sentence in natural language is that the relationship among words in our summary is more global. Due to the fixed time span (i.e., 60 minutes) applied to our summary and the nature of our medical words, we can assume that any words in a summary highly likely to have close interaction with the rest of the words. That is, the window size for training CBOW can be maximized to the whole summary. As introduced in the related work, CBOW projects words into a vector space by predicting an output word given a set of input context words. In our setting, we train an entry-level weight matrix $\boldsymbol{W} \in \mathbb{R}^{|V| \times d}$ where each row of $\boldsymbol{W}$ is a d-dimensional vector representation of the associated input medical word among total $|V|$ vocabulary words, and another matrix $W' \in \mathbb{R}^{d \times |V|}$ for output words. The matrix $\boldsymbol{W}$ can be learned by maximizing the following log-likelihood:

$$\min_{\boldsymbol{W}, \boldsymbol{W'}} {}^1\!/_L \sum_{l=1}^{L} \sum_{i:w_i \in S^l} \log p(w_i | \{w_j | j : w_j \in S^l, j \neq i\}),$$

where $p(w_i | \{w_j | j : w_j \in S^l, j \neq i\}) =$

$$\exp(\boldsymbol{W'}[:,i]^\top \tfrac{1}{n^l - 1} \sum_{j=1, j \neq i}^{n^l} \boldsymbol{W}[j,:]) / \sum_{k=1}^{|V|} \exp(\boldsymbol{W'}[:,k]^\top \tfrac{1}{n^l - 1} \sum_{j=1, j \neq i}^{n^l} \boldsymbol{W}[j,:])$$

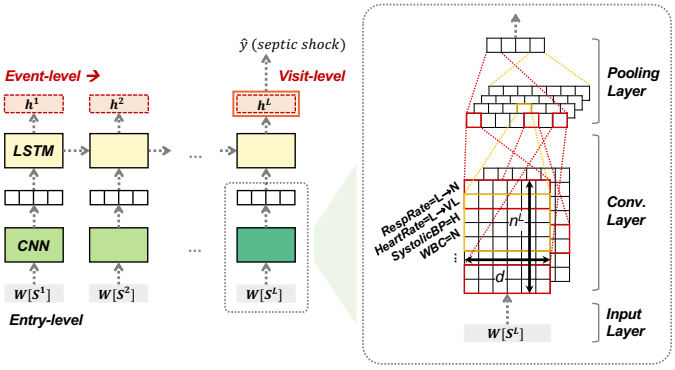

Fig. 2. Model Architecture of *MuLan*

Note that all entry-level vector representations are updated using gradient descent, and we denote $\boldsymbol{W}$ as ***"pre-trained" entry-level embedding***.

C. Event/Visit-level Representation Learning

Different from the entry-level, event/visit-level representations are learned in a supervised way. As shown in Fig. 2, the pre-trained entry-level embedding is used as input and fine-tuned simultaneously by involving knowledge specific to septic shock progression. For this stage, CNN and LSTM are employed to capture intra-event co-occurrence patterns and inter-event temporal dependencies respectively from our temporal summaries and make a prediction.

Multichannel CNN: Motivated by Kim et al. [38], our CNN component consists of three types of layers: *an input layer*, *a convolutional layer*, and *a pooling layer* as shown in Fig. 2. The ***input layer*** prepares an embedding matrix, which is a concatenation of the vector representations corresponding to the n^i words that appear in a i-th sentence S^i. Consequently, $n^i \times d$ matrix is passed to the next layer, where d is the pre-defined dimension size of the entry-level embedding W. In the ***convolutional layer***, a set of filters with different sizes is applied to a set of word vectors in a pre-defined window to generate new features. Here, the filters look for the most informative sets of words to septic shock progression. Especially for septic shock, which doesn't have distinctive single biomarkers, investigating sets of symptoms (i.e., words) is essential to capture the latent mechanism of its progression. We expect that this layer can capture critical sets of words and extract their patterns for the septic shock prediction. Additionally, we employ two channels of input embedding matrix; one embedding matrix is *fine-tuned* during the training process while the other plays a role in preventing over-fitting. The generated features are concatenated to form a set of feature maps at the end. Finally, ***the pooling layer*** extracts the most distinctive feature from each feature map to focus on the strongest co-occurrence patterns. The final output at each time interval is fed into the next LSTM layer.

LSTM: Each cell of LSTM is designed to control information flow across the cells at different timestamps. Therefore, LSTM is managed to model the underlying disease progression mech-

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

anism. The yellow boxes in Fig. 2 indicate the LSTM cells at different time intervals. Each cell returns a hidden state $\mathbf{h^i}, i \in [1, L]$ which is denoted as ***event-level embedding*** since it contains the most salient information up to the corresponding timestamp. The final hidden state $\mathbf{h^L}$, which represents the whole visit, is used to predict the output label y and denoted as ***visit-level embedding***.

Finally, all three levels of embeddings included in this stage are either fine-tuned or updated using gradient descent on the model's cross-entropy loss.

IV. Experiment

A. Dataset Descrption

Our dataset comprises 2.5 years' anonymized EHRs collected from Christiana Care Health System (CCHS) from July 2013 to December 2015. In total, the dataset contains 119,857 patients, 210,289 visits, and 10,412,729 medical events. Among the total population, 52,919 visits and 4,224,567 events are selected as a *study population*, which includes the patients with suspected infection identified by the presence of any type of anti-infective (antibiotic, antiviral, or antifungal) administration, or a positive test result of Point of Care Rapid. The rules for identifying patients with suspected infection and labeling criteria for sepsis stages in the following section were designed by the two leading clinicians with extensive experience on this subject from two leading healthcare systems CCHS and Mayo.

Labeling: International Classification of Diseases (ICD-9) codes are widely used as true labels for various supervised tasks. However, developing a model solely depending on ICD-9 codes can be problematic in that the codes are introduced for administrative purposes and they are assigned at visit-level; previous studies have demonstrated the unreliability of the codes as ground truth [56]. Based on Sepsis-3 definitions [57], our domain experts identified septic shock at event-level as having received vasopressor(s) or having had persistent hypotension for more than an hour (i.e., systolicBP $< 90 mmHg$; or mean arterial pressure $< 65 mmHg$; or drop in systolicBP $> 40 mmHg$ in an 8-hour window)

Sampling: We identified 2,964 shock and 40,513 non-shock visits from the *study population*, which satisfy both ICD-9 and our expert-defined rules. Given the class imbalance, we conducted a stratified random sampling on the non-shock visits while keeping the same distribution of age, gender, ethnicity, length of stay, and the number of records in both shock and non-shock visits. As a result, the final dataset contains 5,928 visits with 795,314 events. We denote these visits as "clean" visits as they are used to evaluate our framework.

Feature Selection: In addition to the identifier, timestamp, and current location, we include four types of features that are used to label sepsis stages; 1) vital signs: heart rate, temperature, systolicBP, etc., 2) lab test results: BUN, lactate, platelet, white blood cell count (WBC), etc., 3) interventions: oxygen source, FiO2, drug administration, intravenous therapy, etc., and 4) assessment results: Glasgow coma scores, Glasgow best verbal response score.

Time Irregularity and Data Sparsity: Each patient's visit within EHRs consists of multivariate events with irregular time intervals. In the study population, the time intervals between two consecutive events vary from *0.94* seconds to *64.38 hours* due to the different measuring frequencies of various features. This causes the data sparsity in EHRs where, on average, more than 80% of values are missing. Particularly, lab results have the highest missing rates which vary from 92.83% to 99.97%.

Pre-training Entry-level Embedding: From the *study population*, we randomly sampled 10,000 visits with 927,131 events that are not included in the "clean" visits, for pre-training. Specifically, 5,000 visits each are sampled from the two classes determined by the expert rule. For shock visits, all events up to the first onset of septic shock are selected while the events in non-shock visits are cut to have the same distribution of length as the shock visits. Again, this "held-off" visits are NOT used for the evaluation.

B. Experimental Setup

Fig. 3 shows the setting for our prediction task: predicting whether a patient will develop *septic shock* ***m*** hours later given the last ***n*** hours of patient's records. To minimize the risk of developing septic shock, m varies from 24 to 48 hours as requested by our clinicians. To accomplish this task, the shock visits are right-aligned by their first onset of septic shock and the non-shock visits by a truncated time point. The non-shock visits are cut to have the same distribution of length as shock visits to prevent possible bias. We focus on five days of patients' records described as *focus window* in Fig. 3 due to the rapid progression of sepsis. Furthermore, we exclude visits with less than two events since such short visits do not provide sufficient information. For training purposes, the model only utilizes n hours of records in the *observation window*. Note that as the size of *hold-off window* increases, the number of patients' visits diminishes significantly: only 41% of visits remains for 24 hours early prediction task and 29.8% for 48 hours. As a result, our task would become more challenging as the number of training visits decreases.

We compared *six* event-level representation models by feeding them directly to an LSTM classifier. The models consist of the four variations of *MuLan* and two baselines: *EHR* for using raw EHRs as input and *Med2Vec*. Note that initially all six models use patients' visit trajectories as input and the main difference among them is different types of representation learning explored. However, in our preliminary experiments, *Med2Vec* performed badly even worse than *EHR*. Therefore,

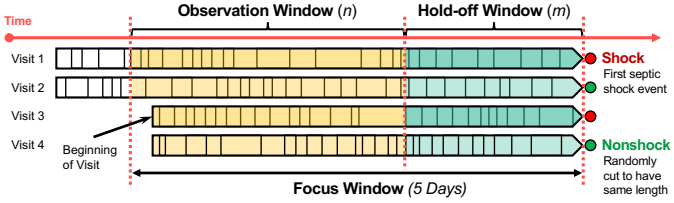

Fig. 3. Right-aligned Setting for Septic Shock Early Prediction

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

TABLE I
PERFORMANCE (± STANDARD DEVIATION) FOR SEPTIC SHOCK EARLY PREDICTION(HOLD-OFF WINDOW = 40 HOURS)

Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
Baseline	*EHR*	0.749(± 0.006)	**0.814**(± 0.016)	0.664(± 0.023)	0.731(± 0.010)	0.818(± 0.004)
	Med2Vec	**0.763**(± 0.007)	0.766(± 0.017)	**0.776***(± 0.027)	**0.770**(± 0.008)	**0.847**(± 0.004)
$MuLan_{sum}$	*TF-IDF*	0.761(± 0.015)	0.792(± 0.008)	**0.726**(± 0.040)	0.757(± 0.022)	0.829(± 0.010)
	Count	**0.768**(± 0.008)	**0.811**(± 0.021)	0.717(± 0.041)	**0.760**(± 0.016)	**0.853**(± 0.006)
$MuLan_{emb}$	*Entry*+	0.769(± 0.017)	0.801(± 0.025)	0.736(± 0.052)	0.765(± 0.024)	0.854(± 0.011)
	Entry.CNN	**0.784***(± 0.012)	**0.820***(± 0.022)	**0.743**(± 0.041)	**0.779***(± 0.017)	**0.860***(± 0.005)

· For each block, the best performing model is in **bold**, and the best model across ALL is underlined and marked with *.

as the original *Med2Vec* in [4] uses patients' demographic information as well as visit trajectories, we include such supplementary information (i.e., age and gender) for *Med2Vec* only to keep its predictive power. It is important to know that age and gender are often closely related to septic progression as the septic shock is more likely to occur among older people than young ones, and male than female [54], [58], [59]. So the question here is: would our proposed *MuLan* variational models using patients' visit trajectories only outperform or perform as good as *Med2Vec* with both types of information.

Two Baselines include:

- ***EHR***: Raw (i.e., not discretized) EHRs with numerous enhancements: 1) standardization with feature-wise mean and standard deviation, 2) aggregation of the events within an hour to an event together with statistical features such as min/max for numerical variables, and 3) expert rule-based imputation, known to be very powerful [60], where vital signs and laboratory results are carried forward with the last value for 8 hours and 24 hours, respectively. The remaining missing values are imputed with feature-wise mean.
- ***Med2Vec***: The final representation of *Med2Vec* [4] which incorporates our summary and patients' demographic information as input. It is noteworthy to mention that for a fair comparison, we changed the classifier from logistic regression to LSTM, which is a more powerful classifier for modeling disease progression.

Four Variations of *MuLan*

Two $MuLan_{sum}$ models representing the generated *summary* vs. Two $MuLan_{emb}$ models from the learned *embedding*.

Two $MuLan_{sum}$ models include:

- ***Count***: A vector representation of a summary that records the number of word occurrences in each time interval.
- ***TF-IDF***: A vector representation of a summary that considers the word importance in each document. The importance level is calculated by an equation $(1+log f_{w,v}) * log(1+\frac{N}{n_w})$, where $f_{w,v}$ is the number of times word w shows up in visit v, N is the number of visits, and n_w is the number of visits that contain the word w.

Two $MuLan_{emb}$ models include:

- ***Entry+*** Sum of the entry-level embeddings in a summary.
- ***Entry*.CNN** Our final model with convolutional filters.

C. Evaluation and Setting

Our evaluation metrics include accuracy, recall, precision, F1-score, and AUC. Accuracy, F1-score, and AUC are widely used in the machine learning domain to measure the prediction performance, while recall and precision are commonly used in the healthcare domain. Among the metrics, we focus on F1-score and AUC as they provide balanced measures. For LSTM, a mini-batch stochastic optimizer with a batch size of 50 and 72 hidden units are employed. The training epochs are set to a maximum of 20 with early stopping. For CNN, a filter size of 50 and a kernel size of $\{3, 5, 7\}$ are used with the ReLU activation function and a dropout rate of 0.8. For *Med2Vec* and *MuLan*, a dimension size of 100 is used for both representations, and a batch size of 250 and 100 are used respectively. For evaluation, we randomly partition the dataset (i.e., "clean" visits) into the training, validation, and testing sets with a ratio of 70%, 15%, and 15%. The performances of the models are reported with the average values of mean and standard deviations of 10 repeated times of experiments with a random model initialization.

V. RESULT AND DISCUSSION

A. Septic Shock Early Prediction (40 Hours Hold-off Window)

Table I shows the result of 40 hours early prediction of septic shock. *EHR* directly uses original EHRs as input and is shown to be a strong baseline with more than 80% of precision and AUC even in this challenging task. It can be explained by the preprocessing steps applied to *EHR*, such as expert rule-based imputation. As expected, *Med2Vec* outperforms *EHR* on all metrics except for the precision, which suggests that representation learning can improve the prediction performance. The biggest improvement is observed in recall (11.2% increase) and followed by AUC (3.9% increase) and F1-score (2.9% increase).

Between the two $MuLan_{sum}$ approaches (2nd section), *Count* outperformed *TF-IDF* across all metrics except for the recall. This indicates that simply knowing how many times a symptom occurs in each event is more critical for septic shock early prediction than using keywords. Furthermore, *Count* significantly outperforms the baseline *EHR* across all metrics, especially it improved recall by 5.3%. This demonstrates the first stage of our framework, EHRs summarization, was effective. Even when compared with *Med2Vec*, *Count* shows

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

TABLE II
OVERALL PERFORMANCE (± STANDARD DEVIATION) FOR SEPTIC SHOCK EARLY PREDICTION (HOLD-OFF WINDOW = 24-48 HOURS)

Type	Model	Accuracy	Precision	Recall	F1-Score	AUC
Baseline	*EHR*	0.750(± 0.011)	**0.796**(± 0.015)	0.685(± 0.021)	0.736(± 0.013)	0.828(± 0.014)
	Med2Vec	**0.773**(± 0.015)	0.777(± 0.014)	**0.776***(±0.025)	**0.776**(± 0.016)	**0.849**(± 0.014)
MuLan$_{sum}$	*TF-IDF*	0.762(± 0.019)	0.789(± 0.014)	0.728(± 0.034)	0.756(± 0.022)	0.835(± 0.019)
	Count	**0.781**(± 0.017)	**0.809**(± 0.013)	**0.747**(± 0.034)	**0.775**(± 0.021)	**0.860**(± 0.016)
MuLan$_{emb}$	*Entry*+	0.779(± 0.018)	0.801(± 0.019)	0.756(± 0.023)	0.776(± 0.018)	0.863(± 0.020)
	Entry.CNN	**0.792***(± 0.017)	**0.818***(± 0.018)	**0.762**(± 0.022)	**0.788***(± 0.017)	**0.865***(± 0.017)

· For each block, the best performing model is in **bold**, and the best model across ALL is underlined and marked with *.

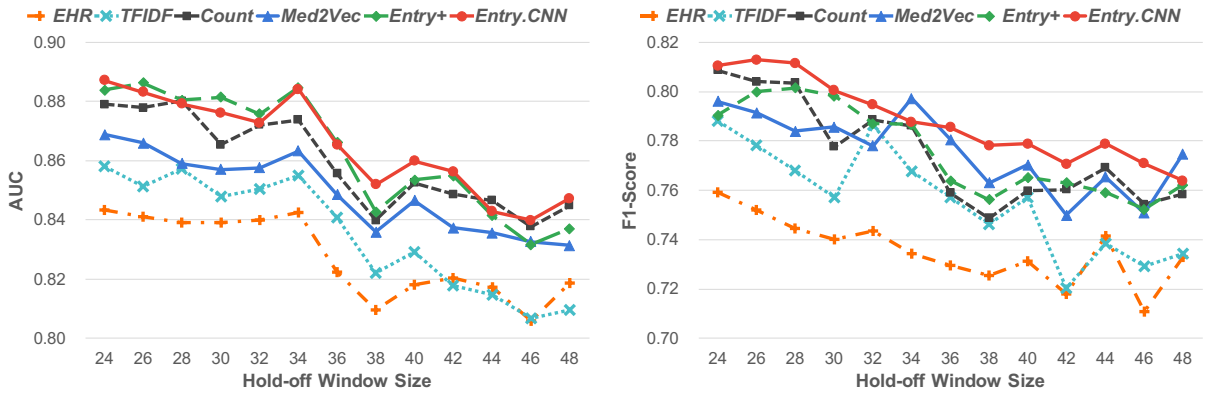

Fig. 4. Overall Performance for Septic Shock Early Prediction (Hold-off Window = 24-48 hours)

comparable performance throughout the metrics even though it did not consider patient demographic information.

Between the two *MuLan*$_{emb}$ models (3rd section), *Entry*.CNN outperforms *Entry*+ across all metrics. The biggest boost is detected in precision (1.9%) followed by accuracy (1.5%). This suggests that *Entry*.CNN's ability to extract informative local patterns for septic shock progression is more effective than *Entry*+'s simple way of generating an input vector. Lastly, by comparing our final model *Entry*.CNN and the best baseline *Med2Vec*, we can observe that our final model outperforms *Med2Vec* in all measures except for recall. One potential explanation is that our *Entry*.CNN did not use patients' demographic information.

B. Overall Septic Shock Early Prediction (24-48 Hours Hold-off Window)

Fig. 4 shows the AUC and F1-scores of different models with varying hold-off window size. Both plots show decreasing trends as the early prediction task becomes more challenging. For AUC, all four proposed models outperform the baseline *EHR* and *Med2Vec* over time except for *TF-IDF*. Particularly, a big improvement is observed between *EHR* and *Count*. In general, two *MuLan*$_{emb}$ models show better performances than *Count* where *Entry*.CNN outperforms *Entry*+ after 38 hours. *Med2Vec*, however, worsens the performance as it is always lower than *Count* and *MuLan*$_{emb}$ models except for 46 hours. For F1-score, all four proposed models outperform *EHR* over time. Similar to AUC, there is a significant enhancement in F1-score between *EHR* and *Count*, which indicates that the summarization of raw format of EHRs into sentences positively affect the performance of the model. Additionally, *Med2Vec* shows comparable performance as *Count* and *Entry*+. Overall, *Entry*.CNN outperforms all other models over time except for 34 and 48 hours where *Med2Vec* performs the best.

Table II shows the overall performance of the models across the given hold-off window sizes. The results show similar patterns found in 40 hours early prediction. *MuLan*$_{emb}$ models generally outperform the two baseline models and *MuLan*$_{sum}$ models. Among all the models, *Entry*.CNN achieves the best performance across all evaluation metrics except for recall, where *Med2Vec* performs the best.

C. Knowledge Discovery from EHRs Summary

One of the most important goals of *MuLan* is about interpretation and visualization. Given that the generated summary contains condensed information of medical data entries in temporal order, it is beneficial to compare the frequent patterns found in shock versus non-shock patients. These patterns can provide supplementary information about sepsis progression, and may help physicians' decision-making. Table III presents the patterns extracted from a pattern mining tool [61] and their corresponding support count over the training population. The numbers in angle brackets indicate relative timestamps, and a set of words followed is symptoms observed at the corresponding time interval. For example, the first row of Table III represents a shock pattern and can be interpreted

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

TABLE III
FREQUENT TEMPORAL PATTERNS FROM EHRS SUMMARIES

Type	Pattern			Support
Shock	⟨0⟩ Location=ED	⟨1⟩ Location=ED	⟨3⟩ SBP=VL, MAP=VL	0.320
Shock	⟨0⟩ Location=ED	⟨2⟩ SBP=VL, DBP=VL, MAP=VL	⟨3⟩ Location=ED	0.222
Shock	⟨0⟩ RespRate=H	⟨2⟩ OxySource=Vent	⟨3⟩ OxySource=Vent	0.219
Non-shock	⟨0⟩ SBP=N, DBP=N, MAP=N	⟨2⟩ Location=NURSE	⟨3⟩ Location=NURSE	0.555
Non-shock	⟨0⟩ RespRate=N, Temp=N	⟨1⟩ Location=NURSE	⟨3⟩ Location=NURSE	0.542

as follows: "A patient was located in an *emergency room* (ED) and stayed for the first two hours (i.e., 0-2). An hour later, he/she developed a symptom of *very low* systolic blood pressure (SBP) and mean arterial pressure (MAP)." The difference between the two patient groups can be found in the compositions of words and support counts. The shock patterns mainly show abnormal conditions while the non-shock patterns contain normal symptoms with higher support counts.

D. Knowledge Discovery from Entry-level Embedding

The entry-level embedding is designed to embrace the heterogeneous relationships between words, namely medical readings. Given that the medical readings indicate the signs produced by a patient's different organ systems (e.g., respiratory system, lymphatic system), the embedding learns the knowledge inherent to the interaction between the organ systems. We present the knowledge discovered in the following subsections.

Pre-trained Entry-level Embedding: To interpret the pre-trained embeddings, we employed three types of word analogy tasks [19] for the relationships between prefixes (e.g., *HeartRate*, *WBC*), suffixes (e.g., *low*, *very high*), and words (e.g., *WBC*=VH).

1) Prefixes (i.e., Features): The relationship between *SystolicBP* (*SBP*) and *DiastolicBP* (*DBP*) is tested with two analogy cases. Note that the plain texts below indicate given analogy questions and the words in bold are the answers from the pre-trained embedding. The answers from both cases demonstrate that our embedding learned the expected relationship as they present *DBP* with the same value as the question.

- (*SBP*=VH : *DBP*=VH) = (*SBP*=VL : ***DBP*=VL**)
- (*SBP*=N→L : *DBP*=N→L) = (*SBP*=L→N : ***DBP*=L→N**)

2) Suffixes (i.e., Values): The below cases suggests that the embedding is able to learn the trend of values. The analogy questions asked increasing trend of the given features (i.e *Temperature*(*Temp*), *SBP*), and we obtained the expected answer: 1) one-level increase (*high* to *very high*) and 2) two-level increase but one-level (*high* to *very high*) since *very high* is the upper limit.

- (*Temp*=L : *Temp*=N) = (*Temp*=H : ***Temp*=VH**)
- (*SBP*=L : *SBP*=H) = (*SBP*=H : ***SBP*= VH**)

3) Words: The closest word to "*OxygenSource*=NotVent→ Vent", which illustrates a specific time period when a patient is being equipped with a ventilator, is chosen to demonstrate whether the functional meaning is captured in the embedding. (Note that ventilators help patients who cannot breathe enough on their own by delivering supplemental oxygen.) We can assume that if additional oxygen is supplied to a patient, the oxygen level of the patient would increase. In our vocabulary, "*PulseOx*" indicates the readings from pulse oximetry that measures the oxygen level in blood. As expected, our result shows that the closest word to "*OxygenSource*=NotVent→Vent" is "*PulsOX*=N→H".

Fine-tuned Entry-level Embedding: To interpret the effect of fine-tuning, 10 most similar words to "*VasopressorTaken*" (see Table IV) are examined. Administration of vasopressor is one of the therapeutic approaches commonly applied to patients with septic shock to maintain their blood pressure [62]. Moreover, our expert rule defines the vasopressor administration as one of the septic shock identification criteria. Therefore, the word "*VasopressorTaken*" is considered as a synonym to *septic shock* in our setting, and any shift in the similar words made by the fine-tuning is investigated.

One interesting finding is that the overall distances towards "*VasopressorTaken*" are greatly diminished. Moreover, the composition of similar words is significantly altered in relation to *septic shock*. For example, in the right table from Table IV, the prefix *Lactate* is ranked at the first and ninth, which is often measured from the patients with suspected *septic shock* [63], [64]. The word "*CurrentLocation*=ED→ICU" are closely related to *septic shock* since *septic shock*-related patients are most likely transferred to intensive care unit (ICU) to get necessary treatment such as administration of vasopressor or fluid resuscitation [65]. The rest of words are very relevant

TABLE IV
MOST SIMILAR WORDS TO "*VassopressorTaken*"

Most Similar Words	Dist.*
HeartRate=N→H	0.648
Bands=N	0.727
RespiratoryRate=L→N	0.745
WBC=H	0.751
Creatinine=H→N	0.755
HeartRate=VL→L	0.759
SedRate=H	0.762
RespiratoryRate=VH→VL	0.765
DiastolicBP=H→VH	0.785
MAP=N	0.791

Most Similar Words	Dist.*
Lactate=N	0.409
AntiFungalTaken	0.409
CurrentLocation=ED→ICU	0.474
BiliRubin=VH	0.495
FIO2=L	0.495
WBC=VH	0.500
Bands=N	0.506
BUN=VH	0.512
Lactate=H	0.527
Creatinine=VH	0.529

* The cosine distances are calculated with the corresponding type of embedding. (Left: Pre-trained, Right: Fine-tuned)

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

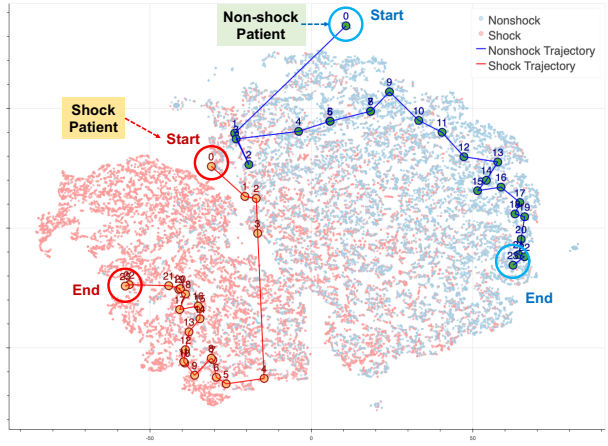

Fig. 5. Visualization of Patient Trajectories. Event-level embeddings from shock patients are marked in red and the opposites are in blue. Among them, two patients' trajectories are displayed with red and blue lines respectively. The number labeled above each point shows the relative order of events.

to the progression of sepsis based on the criteria defined by our clinicians: *AntiFungalTaken* and *Bands*=N are indicators of infection which is the first stage of sepsis; *BiliRubin*=VH (Gastrointestinal), *BUN*=VH (Renal), *Creatinine*=VH (Renal), *GCS*=Moderate (Nervous), and *VerbalGCS*=NotOriented (Nervous) are the criteria of organ dysfunction which is a severe stage that happens prior to septic shock; *WBC*=VH and *SBP*=VH→L are used to diagnose the septic shock. On the other hand, the most similar words from pre-trained embedding are closely related to inflammation, which is the early stage of sepsis.

E. Visualization of Event and Visit-level Embeddings

Among the multilevel representations of *MuLan*, event- and visit-level embeddings can be used to interpret how patients' condition or sepsis progress over time by visualizing the embeddings on a latent vector space. Fig. 5 shows the projection of 1,000 patients' event-level embeddings on 2D scatter plot using t-SNE [66]. One shock and one non-shock patient are sampled to investigate their progression patterns.

The interesting pattern observed is that the two patients are considered to have a similar condition at the early time of events (i.e., the dots are projected in the same area, see 0-2 for red and 1-3 for blue), but they move towards the exact opposite direction. The red dots shift to the leftmost area and the blue dots land at the rightmost area. This suggests that the leftmost area indicates the most critical condition, while the rightmost area can be considered as the healthiest state. This result demonstrates that our framework can visualize patient disease progression in a unified way. In the future, we expect that this could be used to understand new patients' conditions by investigating the location of the projected embeddings.

VI. Conclusion

Modeling patient disease progression is an essential task to support clinical decision making and provide prompt treatment to the patients. In this work, we propose *MuLan*, a Multilevel Language-based representation learning framework, that learns a *hierarchical representation* of EHRs at *entry*, *event*, and *visit* levels. The experimental results on a real-world EHRs demonstrate the predictive power of our model compared to the baselines and the promising ability in interpretation and visualization of our embeddings. In the future, for learning a richer and more personalized representation, we will incorporate patients' demographic information (e.g., age and gender) separately and more details about interventions (e.g., drug names) as they would help discover optimal treatments for patients. Furthermore, we will generate a universal representation of EHRs that can be used in any hospital settings by combining multiple sets of EHRs from different hospitals.

Acknowledgment

This research was supported by the NSF Grants: 1522107, 1651909, 1660878, 1726550, and 2013502.

References

[1] D. Mould, "Models for disease progression: new approaches and uses," *Clin. Pharmacol. Ther.*, vol. 92, no. 1, pp. 125–131, 2012.
[2] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *SIGKDD*. ACM, 2015, pp. 507–516.
[3] B. M. Marlin, D. C. Kale, R. G. Khemani *et al.*, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *IHI*. ACM, 2012, pp. 389–398.
[4] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *SIGKDD*. ACM, 2016, pp. 1495–1504.
[5] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *MLHC*, 2016, pp. 301–318.
[6] J. Zhou, J. Sun, Y. Liu, J. Hu, and J. Ye, "Patient risk prediction model via top-k stability selection," in *SDM*. SIAM, 2013, pp. 55–63.
[7] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NeurIPS*, 2016, pp. 3504–3512.
[8] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *SIGKDD*. ACM, 2017, pp. 787–795.
[9] H. Li, X. Li, X. Jia, M. Ramanathan, and A. Zhang, "Bone disease prediction and phenotype discovery using feature representation over electronic health records," in *BCB*. ACM, 2015, pp. 212–221.
[10] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annu. Rev. Public Health*, vol. 36, pp. 345–359, 2015.
[11] C. Esteban, O. Staeck *et al.*, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," in *ICHI*. IEEE, 2016, pp. 93–101.
[12] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.
[13] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *SIGKDD*. ACM, 2011, pp. 814–822.
[14] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *SDM*. SIAM, 2016, pp. 432–440.
[15] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *MLHC*, 2018, pp. 440–464.
[16] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *MLHC*, 2016, pp. 73–100.
[17] T. Alves, A. Laender, A. Veloso, and N. Ziviani, "Dynamic prediction of icu mortality risk using domain adaptation," in *Big Data*. IEEE, 2018, pp. 1328–1336.

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.

[18] C. Lin, Y. Zhangy, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi, "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm," in *ICHI*. IEEE, 2018, pp. 219–228.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NeurIPS*, 2013, pp. 3111–3119.

[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014, pp. 1188–1196.

[21] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, 2018.

[22] S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *RecSys*. ACM, 2017, pp. 297–305.

[23] S. Kumar, X. Zhang, and J. Leskovec, "Predicting dynamic embedding trajectory in temporal interaction networks," in *SIGKDD*. ACM, 2019, pp. 1269–1278.

[24] M. Sushil, S. Šuster, K. Luyckx, and W. Daelemans, "Patient representation learning and interpretable evaluation using clinical notes," *J. Biomed. Inform.*, vol. 84, pp. 103–113, 2018.

[25] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: a convolutional net for medical records," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 22–30, 2016.

[26] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[27] C. Rhee, T. M. Jones, Y. Hamad, A. Pande, J. Varon *et al.*, "Prevalence, underlying causes, and preventability of sepsis-associated mortality in us acute care hospitals," *JAMA Netw. Open*, vol. 2, no. 2, pp. e187 571–e187 571, 2019.

[28] R. P. Dellinger, M. M. Levy, J. M. Carlet, J. Bion, M. M. Parker, R. Jaeschke, K. Reinhart, D. C. Angus, C. Brun-Buisson, R. Beale *et al.*, "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008," *Intensive Care Med.*, vol. 34, no. 1, pp. 17–60, 2008.

[29] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–1596, 2006.

[30] V. Coba, M. Whitmill, R. Mooney, H. M. Horst, M.-M. Brandt *et al.*, "Resuscitation bundles compliance in severe sepsis and septic shock: improves survival, is better late than never," *J. Intensive Care Med.*, vol. 26, no. 5, pp. 304–313, 2011.

[31] B. Mickiewicz, H. J. Vogel, H. R. Wong, and B. W. Winston, "Metabolomics as a novel approach for early diagnosis of pediatric septic shock and its mortality," *Am. J. Respir. Crit. Care Med.*, vol. 187, no. 9, pp. 967–976, 2013.

[32] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Sci. Transl. Med.*, vol. 7, no. 299, p. 299ra122, 2015.

[33] R. Hotchkiss, L. Moldawer, S. Opal, K. Reinhart, I. Turnbull, and J.-L. Vincent, "Sepsis and septic shock," *Nat. Rev. Dis. Primers*, vol. 2, p. 16045, 06 2016.

[34] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[35] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65 333–65 346, 2018.

[36] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan, and D. Sharp, "E-commerce in your inbox: Product recommendations at scale," in *SIGKDD*. ACM, 2015, pp. 1809–1818.

[37] M. Grbovic and H. Cheng, "Real-time personalization using embeddings for search ranking at airbnb," in *SIGKDD*. ACM, 2018, pp. 311–320.

[38] Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP*. ACL, 2014, pp. 1746–1751.

[39] C. Lin, J. Ivy, and M. Chi, "Multi-layer facial representation learning for early prediction of septic shock," in *Big Data*. IEEE, 2019, pp. 840–849.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[42] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *AAAI*, vol. 34, no. 01, 2020, pp. 833–840.

[43] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "Behrt: transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[44] S. Darabi, M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "Taper: Time-aware patient ehr representation," *IEEE J. Biomed. Health Inform.*, 2020.

[45] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," *arXiv preprint arXiv:1711.03905*, 2017.

[46] Y. Wang, X. Xu, T. Jin, X. Li, G. Xie, and J. Wang, "Inpatient2vec: Medical representation learning for inpatients," in *BIBM*. IEEE, 2019, pp. 1113–1117.

[47] 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 *Big Data*. IEEE, 2017, pp. 1233–1242.

[48] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *ICDM*. IEEE, 2015, pp. 721–726.

[49] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *SIGKDD*. ACM, 2014, pp. 85–94.

[50] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Paient subtyping via time-aware lstm networks," in *SIGKDD*. ACM, 2017, pp. 65–74.

[51] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *SIGKDD*. ACM, 2015, pp. 705–714.

[52] Q. Suo *et al.*, "A multi-task framework for monitoring health conditions via attention-based recurrent neural networks," in *AMIA*, 2017, p. 1665.

[53] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, p. 6085, 2018.

[54] X. Yang, Y. Zhang, and M. Chi, "Time-aware subgroup matrix decomposition: Imputing missing data using forecasting events," in *Big Data*. IEEE, 2018, pp. 1524–1533.

[55] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, "Mining recent temporal patterns for event detection in multivariate time series data," in *SIGKDD*. ACM, 2012, pp. 280–288.

[56] K. K. Giuliano, "Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis," *Am. J. Crit. Care*, vol. 16, no. 2, pp. 122–130, 2007.

[57] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[58] N. Nasir, B. Jamil, S. Siddiqui, N. Talat, F. A. Khan, and R. Hussain, "Mortality in sepsis and its relationship with gender," *Pak. J. Med. Sci.*, vol. 31, no. 5, p. 1201, 2015.

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

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

[61] Y. Hirate and H. Yamana, "Generalized sequential pattern mining with item intervals," *J. Comput.*, vol. 1, no. 3, pp. 51–60, 2006.

[62] T. W. Scheeren, J. Bakker, D. De Backer, D. Annane, P. Asfar, E. C. Boerma, M. Cecconi, A. Dubin, M. W. Dünser, J. Duranteau *et al.*, "Current use of vasopressors in septic shock," *Ann. Intensive Care*, vol. 9, no. 1, p. 20, 2019.

[63] C. Rhee, M. V. Murphy, L. Li, R. Platt, and M. Klompas, "Lactate testing in suspected sepsis: trends and predictors of failure to measure levels," *Crit. Care Med.*, vol. 43, no. 8, p. 1669, 2015.

[64] S. M. Lee and W. S. An, "New clinical criteria for septic shock: serum lactate level as new emerging vital sign," *J. Thorac. Dis.*, vol. 8, no. 7, p. 1388, 2016.

[65] K. Thompson, B. Venkatesh, and S. Finfer, "Sepsis and septic shock: current approaches to management," *Intern. Med. J.*, vol. 49, no. 2, pp. 160–170, 2019.

[66] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Authorized licensed use limited to: University of Illinois. Downloaded on March 05,2022 at 01:58:14 UTC from IEEE Xplore. Restrictions apply.