A Treatment Engine by Predicting Next-Period Prescriptions

Bo Jin
Dalian University of Technology
Dalian, China
jinbo@dlut.edu.cn

Haoyu Yang
Dalian University of Technology
Dalian, China
hao_yu_yang@mail.dlut.edu.cn

Leilei Sun*
Tsinghua University
Beijing, China
sunll@sem.tsinghua.edu.cn

Chuanren Liu
Drexel University
Philadelphia, PA
chuanren.liu@drexel.edu

Yue Qu
Dalian University of Technology
Dalian, China
example15415941@qq.com

Jianing Tong
Tongji University
Shanghai, China
tongjianing@mail.dlut.edu.cn

ABSTRACT

Recent years have witnessed an opportunity for improving healthcare efficiency and quality by mining Electronic Medical Records (EMRs). This paper is aimed at developing a treatment engine, which learns from historical EMR data and provides a patient with next-period prescriptions based on disease conditions, laboratory results, and treatment records of the patient. Importantly, the engine takes consideration of both treatment records and physical examination sequences which are not only heterogeneous and temporal in nature but also often with different record frequencies and lengths. Moreover, the engine also combines static information (e.g., demographics) with the temporal sequences to provide personalized treatment prescriptions to patients. In this regard, a novel Long Short-Term Memory (LSTM) learning framework is proposed to model inter-correlations of different types of medical sequences by connections between hidden neurons. With this framework, we develop three multifaceted LSTM models: Fully Connected Heterogeneous LSTM, Partially Connected Heterogeneous LSTM, and Decomposed Heterogeneous LSTM. The experiments are conducted on two datasets: one is the public MIMIC-III ICU data, and the other comes from several Chinese hospitals. Experimental results reveal the effectiveness of the framework and the three models. The work is deemed important and meaningful for both academia and practitioners in the realm of medical treatment and prediction, as well as in other fields of applications where intelligent decision support becomes pervasive.

CCS CONCEPTS

• **Information systems → Data mining**; • **Applied computing → Health care information systems**;

KEYWORDS

Prescription Prediction; Temporal Sequences; EMRs; Treatment

ACM Reference Format:
Bo Jin, Haoyu Yang, Leilei Sun, Chuanren Liu, Yue Qu, and Jianing Tong. 2018. A Treatment Engine by Predicting Next-Period Prescriptions . In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3220095

1 INTRODUCTION

According to the statistics of the World Bank, the healthcare expenditure of the United States had surpassed 17.0% of its total GDP in 2016[2]. However, most of the nationals are still unsatisfied with the quality, efficiency, and cost of the healthcare services. Christensen et al. [9] attributed this dilemma to the low-efficiency of the current healthcare systems and appealed people to seek for a new design of healthcare services. On the other hand, a large-volume of Electronic Medical Records (EMRs) have been accumulatd due to the wide use of the Hospital Information Systems (HIS) across the world. With the rapid development of data science and artificial intelligence, recent years have witnessed an opportunity of systematically improving healthcare quality and efficiency by mining large-scale EMRs, which attracted remarkable attention of both academia and practitioners [24, 28].

In this paper, we develop a treatment engine to automatically provide data-driven and evidence-based treatment guidelines to patients and doctors. The idea is to predict next-period prescriptions by learning knowledge hidden in large-scale data including demographics, disease conditions, laboratory results, diagnostics, and historical treatment records of patients. By leveraging knowledge learned from the treatment records provided by senior doctors and EMRs of patients with satisfied treatment outcomes, such an engine can not only improve healthcare efficiency but also enhance healthcare quality via guiding all doctors to avoid accidental faults as well as make informed clinical decisions.

However, it is a challenging task to develop such an automatic treatment engine due to the intrinsic complexity of EMR data and the inter-correlations of different kinds of medical factors. EMRs consist of rich temporal and heterogeneous information. Specifically, demographics and admitting diagnosis can be seen as static information during the hospitalization period, while laboratory results are recorded with timestamps. Meanwhile, prescriptions (or doctor orders) have temporal interval labels recording both starting and ending points of time. Additionally, different kinds of medical factors can affect each other dynamically. For example, we should take the historical treatment records, the recent physical

[*] Corresponding Author.

KDD '18, August 19–23, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220095

examination results, and the demographic information (e.g., age, gender) into account to predict the next-period prescriptions. In sum, EMRs contain textual, numerical, categorical, and time data, and the treatment engine should be able to handle multifaceted longitudinal data effectively to learn the hidden treatment patterns.

To address these challenges, our treatment engine uses a novel Long Short-Term Memory (LSTM) learning framework with three heterogeneous LSTM models. In particular, we design a heterogeneous LSTM structure, which contains several hidden temporal sequences, with each hidden temporal sequence capturing the dynamics of one type of medical sequences. In the next-period prescription prediction model, we construct one hidden temporal sequence for modeling prediction sequence and the other hidden temporal sequence for modeling physical examination results. Correspondingly, one hidden sequence reflects the treatment course, and the other hidden sequence reflects the recovery progress. Then, three different heterogeneous LSTM models are developed to explore the interactions of different medical sequences, where the interactions of hidden states are bidirectional and parallel in fully-connected heterogeneous LSTM. The interactions are from hidden physical states to treatment hidden states in partially-connected heterogeneous LSTM, and the physical examination results are directly imposed on treatment hidden states in decomposed LSTM models. Finally, for personalized treatments, two types of static factors are incorporated with the hidden states to predict the next-period prescriptions. One type is the demographics information such as age and gender reflecting the physical conditions of patients; the other type is the diagnostics reflecting the disease severities of patients.

This paper has both practical and theoretical contributions. From the practical perspective, the paper designs a new type of data-driven healthcare service, which has the potential to improve healthcare efficiency and quality by leveraging knowledge learned from a massive amount of EMR data. From the theoretical perspective, the paper studies new LSTM learning structures to handle multifaceted longitudinal data by modeling inter-correlations of different types of temporal sequences. The proposed heterogeneous LSTM models can also be used in other application scenarios, such as workflow management and human behavior analytics.

The rest of the paper is organized as follows. In Section 2, we introduce the formalization and framework. Next, we introduce the proposed heterogeneous LSTM models in Section 3, and evaluate them with real-world data in Section 4. We discuss the related work in Section 5 and conclude our work and highlight future research directions in Section 6.

2 FORMALIZATION AND FRAMEWORK

In EMRs, a patient's hospitalization information is almost completely stored, which includes all the prescriptions and physical examination indicators. Additionally, the demographics and diagnostics are also recorded when a patient visits a hospital. A unique patient ID associates this information. This paper studies how to predict the next-period prescriptions, where all the above information will be taken into account. Figure 1 illustrates the EMR data used in our research.

Demographic information: Demographic information includes the age, gender, address, ethnicity, education, and other information of a patient. This piece of information plays an important role in clinical decisions, e.g., therapeutic regimen design and dosage selection. During a patient's hospitalization, demographic information can be seen to be static, which is denoted by $\mathbf{x}^{demo}$.

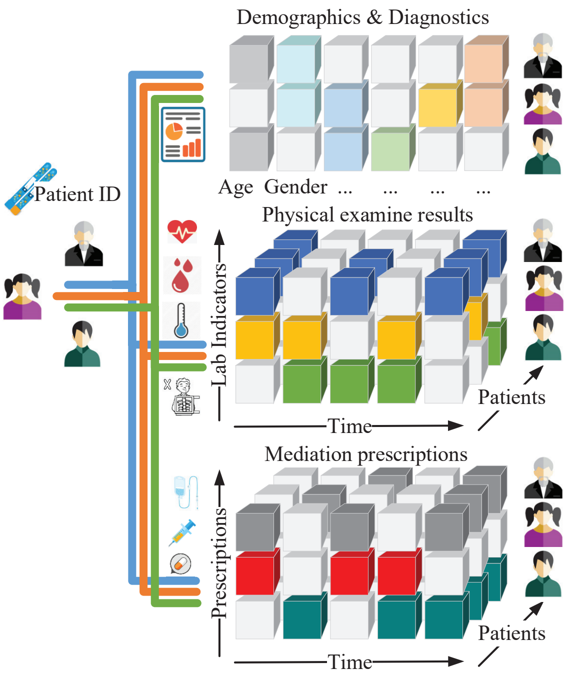

Figure 1: Four Categories of EMRs Data Used in Our Treatment Engine.

Diagnostic information: Diagnostic information includes the disease type, severity, and complications. The primary diagnostic results are given before the treatment process, which are fixed during the hospitalization. We use $\mathbf{x}^{diag}$ to represent diagnostic information, and $\mathbf{x}^s = [\mathbf{x}^{demo}\ \mathbf{x}^{diag}]$ to indicate static information during hospitalization.

Physical examination results: To evaluate the treatment effect or to monitor the recovery progress, a patient may be examined several times during a hospitalization. We use $\mathbf{x}_t^l$ to represent the physical examination result at time t, where $x_{j,t}^l$ is the value of the j-th examined indicator at time t.

Medication prescriptions: A prescription includes the medicine name, delivery route, daily dosage, starting time, and ending time. We use $\mathbf{x}_t^m$ to indicate the active prescriptions at time t, where $x_{j,t}^m = 1$ if medicine j is used in t-th day; $x_{j,t}^m = 0$, otherwise. The treatment sequence $\{\mathbf{x}_1^m, \mathbf{x}_2^m, \cdots, \mathbf{x}_t^m, \cdots\}$ includes the prescription records from the first day to the last day.

Figure 2 is the schematic of the treatment engine proposed in this paper, it can be seen that such a predictor should be able to: 1) handle longitude temporal sequences with variable length; 2) capture the dynamics of multifaceted temporal sequences simultaneously; 3) model the interactions of multiple heterogeneous temporal sequences; and 4) incorporate static information with dynamic information. These give rise to the difficulties for the design

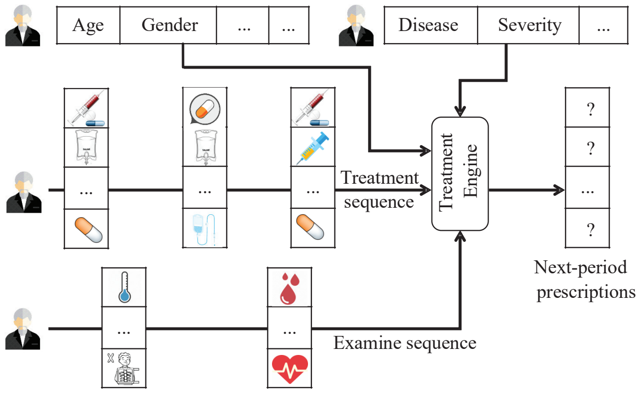

Figure 2: The Schematic of Treatment Engine.

of treatment engine, which motivates us to develop a novel LSTM learning framework for predicting the next-period prescriptions.

The aim of this paper is to train a machine learning model $f(\cdot, \cdot, \cdot)$, which can predict the next medication prescriptions for a patient according to the patient's demographic and diagnostic information, medication history and physical examine indicators. That is,

$$\mathbf{y}_t = f(\mathbf{x}^s, \mathbf{h}^m_{t-1}, \mathbf{h}^l_{t-1}), \tag{1}$$

where $\mathbf{h}^m_{t-1}$ is an aggregation of historical medications from $\mathbf{x}^m_1$ to $\mathbf{x}^m_{t-1}$, and $\mathbf{h}^l_{t-1}$ is an aggregation of all available physical examination results.

3 METHODOLOGY

To realize the treatment engine, we propose a novel long-short memory learning framework in this paper and construct three multifaceted heterogeneous LSTM models. This section discusses the details of the proposed models.

3.1 Recurrent Neural Network

Recurrent neural network (RNN) was designed particularly to deal with temporal data, which is a neural network with chain-like structure. Each hidden unit of RNN receives the previous unit's message and gives its own message to the next unit. Benefiting from this structure, RNN can capture the dynamics underlying temporal sequences. LSTM[15] is a particular RNN, which was proposed to overcome the vanishing and exploding gradient problems of RNN[14]. Due to its ability to learn both long and short dependencies of temporal dynamics, it has been widely studied and used in recent years. In this paper, we develop our treatment engine based on LSTM models.

The basic LSTM model takes a temporal sequence $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t\}$ as input. In step t, it updates the hidden state $\mathbf{h}_t$ by combining the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. However, the prediction of next-period prescriptions needs to model multifaceted heterogeneous temporal sequences, and to combine static information with sequential dynamics. Therefore, the basic LSTM model cannot be used any longer, we propose three novel heterogeneous LSTM models as follows.

3.2 Fully Connected Heterogeneous LSTM

Here, the learning model should take multiple temporal sequences into account simultaneously. We will propose three types of heterogeneous LSTM to deal with multiple temporal sequences. The first model is the Fully Connected Heterogeneous LSTM (LSTM-FC). In this model, all of the heterogeneous sequences are taken as input, and sequential hidden states are constructed for each temporal sequence. The mathematical expressions of LSTM-FC are as follows:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{x}^m_t, \mathbf{x}^l_t, \mathbf{h}_{t-1}] + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{x}^m_t, \mathbf{x}^l_t, \mathbf{h}_{t-1}] + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{x}^m_t, \mathbf{x}^l_t, \mathbf{h}_{t-1}] + \mathbf{b}_o) \\
\tilde{\mathbf{C}}_t &= \tanh(\mathbf{W}_c[\mathbf{x}^m_t, \mathbf{x}^l_t, \mathbf{h}_{t-1}] + \mathbf{b}_c) \\
\mathbf{C}_t &= \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{C}}_t \\
\mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{C}_t)
\end{aligned}$$

The structure of the fully connected heterogeneous LSTM unit is shown in Figure 3.

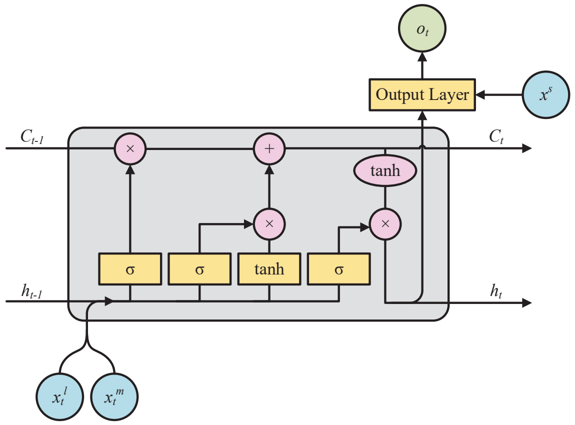

Figure 3: Structure of Fully Connected LSTM.

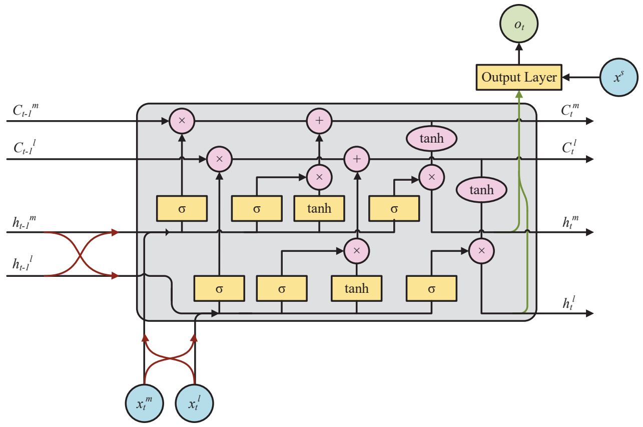

Figure 4: Structure of Fully Connected LSTM in Dual Mode.

Furthermore, we can see that the fully connected LSTM model can be divided into separate hidden sequences when we denote $\mathbf{h}_t$ as

$$\mathbf{h}_t = [\mathbf{h}_t^m, \mathbf{h}_t^l] \tag{2}$$

Let $\mathbf{W}_*$ denote the weight of different LSTM gates, and $\mathbf{b}_*$ denotes the bias. Then we can rewrite the fully connected LSTM model as

$$\boldsymbol{h}_t = f(\mathbf{W}_*[\mathbf{x}_t^m, \mathbf{x}_t^l, \mathbf{h}_{t-1}] + \mathbf{b}_*)$$

$$\begin{bmatrix} \mathbf{h}_t^m \\ \mathbf{h}_t^l \end{bmatrix} = f(\begin{bmatrix} \mathbf{W}_*^m \\ \mathbf{W}_*^l \end{bmatrix} \begin{bmatrix} \mathbf{x}_t^m \\ \mathbf{x}_t^l \\ \mathbf{h}_{t-1}^m \\ \mathbf{h}_{t-1}^l \end{bmatrix} + \begin{bmatrix} \mathbf{b}_*^m \\ \mathbf{b}_*^l \end{bmatrix}) \tag{3}$$

where f denotes the combination of all the gate functions and cell state functions to update the hidden state. We can get two hidden states by matrix block multiplication from Equation (3):

$$\begin{aligned} \mathbf{h}_t^m &= f(\mathbf{W}_*^m[\mathbf{x}_t^m, \mathbf{x}_t^l, \mathbf{h}_{t-1}^m, \mathbf{h}_{t-1}^l] + \mathbf{b}_*^m) \\ \mathbf{h}_t^l &= f(\mathbf{W}_*^l[\mathbf{x}_t^m, \mathbf{x}_t^l, \mathbf{h}_{t-1}^m, \mathbf{h}_{t-1}^l] + \mathbf{b}_*^l) \end{aligned} \tag{4}$$

We can infer from Equation (4) that the predicting sequence $\mathbf{x}_t^m$ can directly change the hidden state of additional sequence $\mathbf{h}_t^l$, and vise versa. This formalization can be seen as the dual mode of the fully connected LSTM. The structure of the dual model is shown in Figure 4. Based on the dual mode, the limitation of the fully connected LSTM can be easily dealt with. In this structure, two heterogeneous input sequences are concatenated with each other (red lines), and the hidden state used to compute the output are actually generated by concatenating two hidden states (green lines). It means that the two input sequences $\mathbf{x}^m$ and $\mathbf{x}^l$ do not have independent hidden states, and the two hidden states contribute equally to the output, which indicates that the inter-corrections of the heterogeneous inputs are not fully utilized.

3.3 Partly Connected Heterogeneous LSTM

As discussed above, fully connected hidden neurons of different temporal sequences may confuse the inherent dynamics of each temporal sequences. To enable flexible interactions of multifaceted temporal sequences, we encode relatively independent memory in each hidden states. Thus, as shown in Figure 5, we propose the Partly Connected LSTM (LSTM-PC) structure as:

$$\begin{aligned} \mathbf{h}_t^m &= f(\mathbf{W}_*^m[\mathbf{x}_t^m, \mathbf{h}_{t-1}^m, \mathbf{h}_{t-1}^l] + \mathbf{b}_*^m) \\ \mathbf{h}_t^l &= f(\mathbf{W}_*^l[\mathbf{x}_t^d, \mathbf{h}_{t-1}^m, \mathbf{h}_{t-1}^l] + \mathbf{b}_*^l) \end{aligned} \tag{5}$$

Compared with fully connected LSTM, the output of partly connected LSTM is only computed by the hidden state $\mathbf{h}^m$. Two input sequences are no longer concatenated to compute hidden state. In this structure, the interactivity between heterogeneous sequences only appears in the hidden state level. To be more specific, $\mathbf{h}_t^m$ is only computed by $\mathbf{x}_t^m$, $\mathbf{h}_{t-1}^m$ and $\mathbf{h}_{t-1}^l$, rather than dependent on $\mathbf{x}_t^l$.

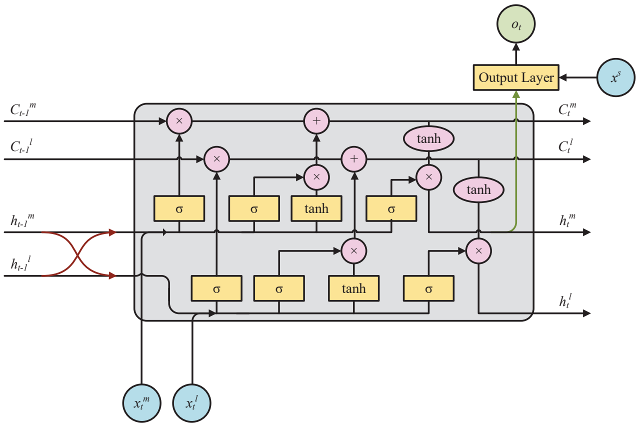

Figure 5: Structure of Partly Connected LSTM.

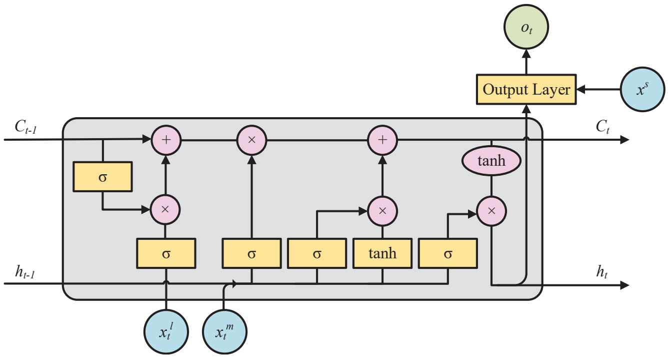

Figure 6: Structure of Decomposed LSTM.

3.4 Decomposed Heterogeneous LSTM

In partly connected LSTM model, we only construct the connections of hidden states between different temporal sequences. In a further step, we explore an even sparse type of connections of hidden neurons. Inspired by [3], we design a novel structure that can take an additional sequence as input. Compared with the partly connected LSTM, the additional hidden state was dropped, only one hidden state for prediction sequence was in this structure. The additional sequence only affects cell state through a unique structure called decomposition gate. The mathematical expressions are listed as follows. For simplicity, we use $\mathbf{g}$ to denote three gates (forget, input and output gates) in LSTM, and $\mathbf{W}_g$, $\mathbf{b}_g$ to denote the weight and bias.

$$\begin{aligned} \mathbf{g}_t &= \sigma(\mathbf{W}_g[\mathbf{x}_t^m, \mathbf{h}_{t-1}] + \mathbf{b}_g) \\ \mathbf{d}_t &= \sigma(\mathbf{W}_{decomp}\mathbf{C}_{t-1} + \mathbf{b}_{decomp}) \\ \tilde{\mathbf{C}}_t^l &= \mathbf{d}_t * \sigma(\mathbf{W}_l\mathbf{x}_t^l + \mathbf{b}_l) \\ \tilde{\mathbf{C}}_t &= \tanh(\mathbf{W}_c[\mathbf{x}_t^m, \mathbf{h}_{t-1}] + \mathbf{b}_c) \\ \mathbf{C}_t &= \mathbf{f}_t * (\mathbf{C}_{t-1} + \tilde{\mathbf{C}}_t^l) + \mathbf{i}_t * \tilde{\mathbf{C}}_t \\ \mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{C}_t) \end{aligned}$$

In this structure, only memories related with prescriptions are preserved. The additional temporal sequences are imposed on the cell state under the control of the previous memory. So we construct a decomposition gate $\mathbf{d}_t$ using previous cell state $\mathbf{C}_{t-1}$, which is used to control the amount of the added information. Controlled by the decomposition gate, we add the additional candidate values $\tilde{\mathbf{C}}_t^l$ to the cell state. In practice, it doesn't make much difference where to add the additional information $\tilde{\mathbf{C}}_t^l$ along the cell state. Using $\mathbf{C}_t = \mathbf{f}_t * (\mathbf{C}_{t-1} + \tilde{\mathbf{C}}_t^l) + \mathbf{i}_t * \tilde{\mathbf{C}}_t$ and $\mathbf{C}_t = \mathbf{f}_t * \mathbf{C}_{t-1} + \tilde{\mathbf{C}}_t^l + \mathbf{i}_t * \tilde{\mathbf{C}}_t$ will get a close result.

3.5 Addition of Static Information

Apart from temporal data, static data of patients are recorded by the hospital information systems. This kind of data generally consists of gender, age, $\cdots$, and other demographic information, which play an important role in personalized treatments. Therefore, we add a fully connected layer to incooperate static facts $\mathbf{x}^s$ with hidden state $\mathbf{h}_t$. For all the three proposed heterogeneous LSTM models, we add the following dense layer to get the output.

$$\mathbf{D} = relu(\mathbf{W}_{dense}\mathbf{h}_t + \mathbf{W}_{static}\mathbf{x}^s + \mathbf{b}_{dense}) \tag{6}$$

$$output = \sigma(\mathbf{W}_{out}\mathbf{D} + \mathbf{b}_{out}) \tag{7}$$

4 EXPERIMENTS

In this section, we evaluate our methods via the experiments on two real-world clinical datasets. The basic LSTM model is used to provide baseline performance. Since the basic LSTM model could not deal with multiple temporal sequences, nor with dynamic and static information simultaneously, only the treatment sequences are used to train and test the basic LSTM model. All these four LSTM models are implemented in Tensorflow with mini-batch stochastic Adam optimizer. For evaluation, we use the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) to assess model discrimination.

4.1 Data Preparation

We evaluate the performance of the proposed methods on two real-world datasets: MIMIC-III and an EMRs Dataset.

4.1.1 *MIMIC-III.*

MIMIC-III[17] is a publicly available large-scale dataset. The dataset contains tables of data related to patients who stayed within the intensive care units (ICU) at Beth Israel Deaconess Medical Center. In this paper, we only consider the subsets in which the patients are diagnosed with the same disease. First, we extract two temporal sequences from the dataset: chart events and prescriptions. Secondary, we extract the patient's demographics from the other tables. Then, the prescription sequence which can be seen as the treatment of patients is selected as the prediction sequence. Finally, the lab event sequence in chart format is selected as an additional sequence.

- **Diagnosis:** For patients with different diseases, the treatment processes are quite different. Thus, we use the patient's diagnosis in ADMISSIONS, the original data table, to choose subsets with the most patients from MIMIC. Thus, there are 11 different diagnoses in total as shown in Tables 1 and 2.
- **Chart Events:** Chart event information is extracted from the table CHARTEVENT. The electronic chart contains much additional information relevant to patients' care: ventilator settings, laboratory indicators, code status, mental status, and so on. We encode all the chart events which the patient receives at same the record time into a feature vector, each entry of which is the value of the event. The event whose occurrence frequency is lower than 100 was dropped.
- **Prescriptions:** Prescriptions information is extracted from the table PRESCRIPTIONS. We use the names of columns DRUG_NAME_POE to identify a specific drug. Similarly to lab events, the events with occurrence frequency lower than 100 are removed.
- **Demographics:** Demographic information is generated from multiple data tables. Insurance, language, religion, marital status and ethnicity of patients are obtained from table ADMISSIONS. The weight of patients is obtained from table INPUTEVENTS_MV. The gender and age of patients are from the data table of PATIENTS. As table PATIENTS only provide the dates of birth, and a patient might have several readmission events at different years, we select the first admission events to record patients' ages, to avoid changes of them.

4.1.2 *EMRs Dataset.*

The EMRs data used in this paper are collected from information systems of 5 public hospitals in Liaoning Province, China. As the goal of this paper is to develop an automatic treatment robot according to large-scale EMRs, we first select a target disease, then pick out all the hospitalization records of the patients with this disease. In this paper, cerebral infarction is selected as target disease, which is a common disease in China. Furthermore, for most of the patients, surgery is not required, so the prescriptions (doctor orders) of a patient record the complete treatment process. In addition to the prescriptions, demographics, diagnostics and laboratory indicators are also collected and stored, which provide both static and dynamic information for the development of an automatic treatment robot.

After collecting the EMRs, clinical doctors do some preprocessing work such as removing fake values, filling the missed values et al. Finally, there are more than 15,000 patients with complete demographic and diagnostic information, each patient with at least one doctor order and laboratory results of one examination. The number of associated doctor orders is larger than 500,000, about 33 doctor orders for each patient.

The collected doctor orders have more than 1000 medicines, some of them are used to treat other diseases as most of the cerebral infarction patients are elderly people hence they might suffer from multiple diseases at the same time, there are also some dietary supplements not closely related with the treatment of cerebral infarction. So we select 132 medicines that are most relevant to cerebral infarction to conduct the experiments.

4.1.3 *Data Preparation.*

The clinic events in both data sets are recorded in minutes. We re-sample them by day to build the physical examination sequences $\mathbf{x}^l$ and medication prescription sequences $\mathbf{x}^m$. The recording time of these heterogeneous sequences could be variable. Even for a single patient, clinical events from different sources might also have different recording time. For example, a patient does some lab examination one day, but the result is measured on the second day, and then the patient obtains his prescription valid. Although these clinical events are recorded on different days, they tie to each other tightly. To make these sequences in synchronization with the same time, we choose a sequence as the base sequence, while the others as the synergic sequence. Time alignment is used between base sequence and synergic sequence, which means that the time step of all the sequences would keep aligned with the base sequence after time alignment. Then, we build the synergic sequence by matching their most recent event with base sequence. In detail, if we use $\{a_2, a_5, a_{10}, a_{17}\}$ to represent the base sequence, $\{b_1, b_{10}, b_{14}\}$ to represent the synergic sequence, the subscript of which denotes the time. The sequence we obtain after time alignment should be $\{a_2, a_5, a_{10}, a_{17}\}$ and $\{b_1, b_1, b_{10}, b_{14}\}$. In this method, two adjacent events in the base sequence may correspond to the same event in synergic sequence (both a_2 and a_5 correspond to b_1). The prescription will be made based on the previous clinical test result, which is quite meaningful in real medical treatments.

Considering the diverse recording time of these heterogeneous sequences, we generate training and test batches using the sequences with the same length, instead of padding the original sequences to a fixed length. In order to acquire better generalization ability, the sequences which are too short or too long are dropped.

4.2 Results Summary

4.2.1 Experiment on MIMIC-III dataset.

Here, we randomly split the data of patients with the same diagnosis into training data (80%) and testing data (20%). We conduct the random splitting process five times and report the average performance of each model. Table 1 compares the average AUROC of the baseline model LSTM and our models, namely, Fully Connected LSTM (LSTM-FC), Partly Connected LSTM (LSTM-PC) and DEcomposed LSTM (LSTM-DE). Table 2 compares the average AUPR of them. All of the models use the same network settings and parameters for comparison. Specifically, the learning rate is set to be 1×10^{-2}, and the dimension of hidden state and fully connected layer are set to be 512 and 256. L2 norm regularization is used with the parameter set to 1×10^{-5}.

First, from the table, we can see that the LSTM-DE outperforms other models in most diagnoses. Second, the performance of LSTM-PC is higher than LSTM-FC in most cases, which indicates that the result computed from some selected hidden states would be better. Finally, the average values of AUROC and AUPR of LSTM-FC are lower than that of LSTM, while the results of LSTM-PC and LSTM-DE are close to LSTM in CORONARY ARTERY BYPASS GRAFT and UPPER GI BLEED. This indicates that the sequence extracted from chart events might not have a close correlation with the predicting prescription sequence. Thus, the performance of the LSTM model without additional information should be better.

After adding low correlated additional information, LSTM-FC would perform worse, while LSTM-PC and LSTM-DE keep the excellent performance. This result can be interpreted by their different unit structures. The output of LSTM-FC is computed by concatenating $\mathbf{h}^m$ and $\mathbf{h}^l$ as shown with the two green lines in Figure 4. The two hidden states have the equal contribution to output. Thus, the ability to discriminate the useful information only rely on the weights in the output layers. The output of LSTM-PC is computed only by $\mathbf{h}^m$ directly, as shown with the green line in Figure 5. The memories in $\mathbf{h}^1$ contribute to the output through more gate functions, which promote the discriminate ability of additional information. In LSTM-DE, decomposition gate is designed specially to control how much additional information should be added into the cell state. It's equivalent to that no additional information is used when the gate is closed (all zeros). In summary, LSTM-DE had the highest control power of additional information, and then LSTM-PC. LSTM-FC have the lowest control power.

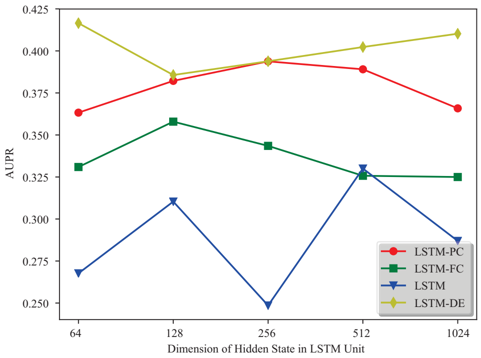

Figure 7: Values of AUROC per Dimension of Hidden State in LSTM Unit on EMRs Test Dataset .

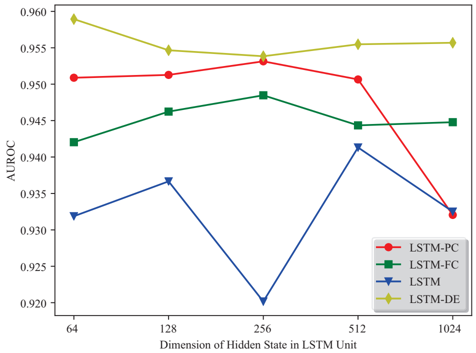

Figure 8: Values of AUPR per Dimension of Hidden State in LSTM Unit on EMRs Test Dataset.

4.2.2 Experiment on EMRs Dataset. Here, we randomly split the data of patients with the same diagnosis into training data (80%) and testing data (20%). Each of the four models is trained for 100 epochs. To avoid exploding gradients, we add L2 norm to the cost function with the parameter of 1×10^{-6}. The other hyper-parameter is tuned to the best. Moreover, we pick up the patients who have at least three records. Then the amount of the selected patients is 19,188.

Table 1: Performance Comparison of LSTM, LSTM-FC, LSTM-PC, LSTM-DE using AUROC on MIMIC-III Dataset

Subsets of MIMIC-III	LSTM	LSTM-FC	LSTM-PC	LSTM-DE
CORONARY ARTERY BYPASS GRAFT	0.8610 ± 0.0215	0.8274 ± 0.0163	0.8578 ± 0.0228	**0.8789 ± 0.0168**
UPPER GI BLEED	0.6656 ± 0.0190	0.5780 ± 0.0516	0.6643 ± 0.0142	**0.6671 ± 0.0089**
CHEST PAIN	0.7715 ± 0.0226	0.7349 ± 0.0305	0.7348 ± 0.0185	**0.7954 ± 0.0126**
ALTERED MENTAL STATUS	0.6629 ± 0.0262	0.6462 ± 0.0386	0.6317 ± 0.0164	**0.6929 ± 0.0186**
ABDOMINAL PAIN	0.6408 ± 0.0125	0.6220 ± 0.0192	0.6958 ± 0.0319	**0.7119 ± 0.0180**
CORONARY ARTERY DISEASE	0.8392 ± 0.0133	0.8045 ± 0.0209	0.8154 ± 0.0329	**0.8795 ± 0.0129**
INTRACRANIAL HEMORRHAGE	0.6305 ± 0.0295	0.6344 ± 0.0316	0.6524 ± 0.0187	**0.6789 ± 0.0091**
CONGESTIVE HEART FAILURE	0.7020 ± 0.0263	0.7260 ± 0.0119	0.7211 ± 0.0199	**0.7567 ± 0.0183**
GASTROINTESTINAL BLEED	0.5849 ± 0.0338	0.5962 ± 0.0301	0.6037 ± 0.0341	**0.6331 ± 0.0344**
SEPSIS	0.6645 ± 0.0115	0.6691 ± 0.0197	0.6730 ± 0.0156	**0.7086 ± 0.0143**
PNEUMONIA	0.7070 ± 0.0098	0.7303 ± 0.0108	0.7346 ± 0.0106	**0.7578 ± 0.0113**

Table 2: Performance Comparison of LSTM, LSTM-FC, LSTM-PC, LSTM-DE using AUPR on MIMIC-III Dataset

Subsets of MIMIC-III	LSTM	LSTM-FC	LSTM-PC	LSTM-DE
CORONARY ARTERY BYPASS GRAFT	0.7749 ± 0.0146	0.7466 ± 0.0501	0.7680 ± 0.0505	**0.8002 ± 0.0467**
UPPER GI BLEED	0.3051 ± 0.0295	0.2520 ± 0.0437	**0.3307 ± 0.0359**	0.3061 ± 0.0318
CHEST PAIN	0.5529 ± 0.0288	0.4972 ± 0.0284	0.4987 ± 0.0342	**0.5426 ± 0.0574**
ALTERED MENTAL STATUS	0.2022 ± 0.0064	0.1744 ± 0.0327	0.1818 ± 0.0240	**0.2238 ± 0.0500**
ABDOMINAL PAIN	0.1850 ± 0.0119	0.1688 ± 0.0180	0.2263 ± 0.0293	**0.2320 ± 0.0174**
CORONARY ARTERY DISEASE	0.7142 ± 0.0346	0.6475 ± 0.0345	0.6705 ± 0.0588	**0.7496 ± 0.0261**
INTRACRANIAL HEMORRHAGE	0.1596 ± 0.0199	0.1570 ± 0.0089	0.1714 ± 0.0101	**0.1878 ± 0.0204**
CONGESTIVE HEART FAILURE	0.2525 ± 0.0201	0.2559 ± 0.0351	0.2740 ± 0.0418	**0.2949 ± 0.0300**
GASTROINTESTINAL BLEED	0.2293 ± 0.0342	0.2344 ± 0.0344	0.2294 ± 0.0279	**0.2310 ± 0.0258**
SEPSIS	0.1650 ± 0.0045	0.1407 ± 0.0109	0.1629 ± 0.0106	**0.1641 ± 0.0069**
PNEUMONIA	0.1527 ± 0.0148	0.1527 ± 0.0073	0.1755 ± 0.0179	**0.1809 ± 0.0134**

Since the dimension of a fully connected layer (output layer) is added to all the three heterogeneous models, we set it to be a fixed number of 512. As shown in Figures 7 and 8, we compare the values of AUROC and AUPR per dimension of hidden state. The value of two metrics in the test dataset is optimized after 100 epochs. In particular, LSTM-DE outperformed other models in all of the dimensions.

Then, we can choose the best dimension of hidden state for all of the four models according to Figures 7 and 8. The curve of two metrics (AUROC and AUPR) along with all 100 epochs in the test dataset are shown in Figures 9 and Figure 10. As shown in these figures, for AUROC, although the results of four methods are similar after 100 epochs, convergence rates of them are quite different. For AUPR, the difference of four models is even more significant, as shown in Figure 10. Compared with AUROC, higher AUPR score indicates that the prescription used for prediction is more likely to be adopted because the positive class is less than the negative class.

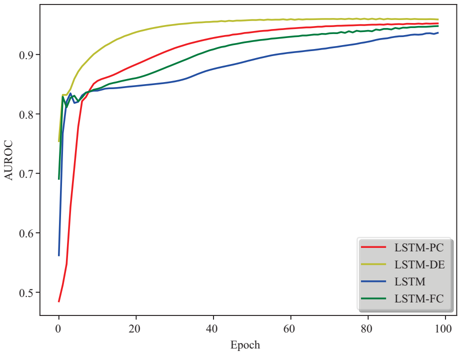

Figure 9: Performance Comparison using AUROC on EMRs Dataset. The dimension of Hidden State in LSTM, LSTM-FC, LSTM-PC, LSTM-DE model are set to be 512, 128, 256, 64.

5 RELATED WORK

Healthcare informatics is a hot topic in the area of data mining. In this section, we provide a brief review of the related work, categorized into two groups.

5.1 Healthcare Data Mining

For assist diagnosis, a similarity-based method was used to design an anomalous tumor motion detector [1], which is called k-weighted

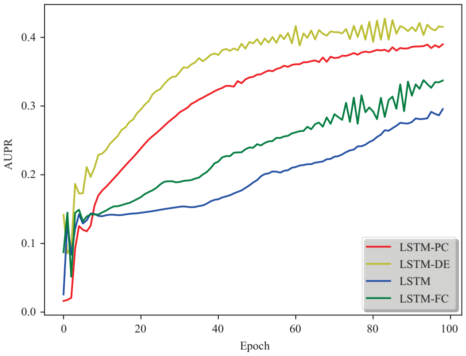

Figure 10: Performance Comparison using AUPR on EMRs Dataset. The dimension of Hidden State in LSTM, LSTM-FC, LSTM-PC, LSTM-DE model are set to be 512, 128, 256, 64.

angular similarity based on regional segmentation. This could identify the different breathing styles in consideration, which was considered suitable for the treatment of the thoracic and abdominal tumor. Chaurasia [4] presented a diagnosis system for detecting breast cancer based on RepTree, RBF Network and Simple Logistic. The correct classification rate of the system was 74.5%. And this system could also be used to diagnose other diseases. Ross [25] used a novel Bayesian nonparametric model base on the concept of disease trajectories for disease subtype identification.This model identified nine subtypes which showed significant associations to seven single nucleotide polymorphisms (SNPs) known to associate with COPD from chest CT scans.

For subtyping disease progress, a Bayesian-based method was introduced [18] in order to improve EMR-based phenotyping by bridging the separated methods, where a nonparametric content-based Poisson factorization was used with better performances in predicting the risk scores than matrix factorization and topic modeling methods. [5] used neural networks which were trained on windows of multivariate clinical time series to discover physiologic patterns and predict health conditions.

For outcome prediction, multi-task learning was used by some researchers to predict the task which will benefit from the data across the population. A multi-task learning regression framework was used to predict PD rating scales of Parkinson's progression [10]. Multi-task learning combined with domain adaptation approaches was built to learn a customized model for each person, in order to predict a person's mood [16]. Transfer learning method was introduced to solve the cross-people activity recognition problem in the field of mobile healthcare [31], which was integrated by decision tree and k-means clustering.

For analyzing treatment process, Sun et al. proposed a data-driven automatic treatment regimen development and extraction method, which could extract treatment regimens from a large number of collected historical treatment seqences, and also generate the most suitable treatment regimen for a patient according to physical conditions and disease severity [26]. Similarly, Yang et al. presented a data-driven framework for standarizing process of medical operations [29]. Such a medical guideline system could help doctors avoid medical accidents.

5.2 Predicting Temporal Events

The prediction of temporal events has been studied in many fields. Basket recommendation and career move predictions are the hot topics in business[30] and workflow[19] fields. Predicting these temporal events, learning possible representations for different types of entities (users, careers) is the key idea. In healthcare fields, temporal events is also a critical factor when extracting knowledge from large-scale EMRs data. Many researchers design different models to mining the possible temporal patterns. EMRs generally aggregates data from many sources, which is very challenging to analyze. Some researchers try to transform the raw data into symbolic time intervals series, remaining clinically meaningful interpretation [7, 22].

Diagnose event prediction is one among the most important in predicting healthcare temporal events based on EMRs. Vasiljeva [27] used machine learning to provide predictions on future diagnoses such as to be experienced by a particular individual, based on the person's existing diagnostic history. Prakash [23] designed a memory network to predict diagnosis given a clinical scenario. Instead of the structured data, they used the raw text extracted from the EMRs as the input. [13] used an expressive data mapping to discover cyclic rules that integrated multiple medical aspects. Using temporal constraints, cyclic patterns discovered from multiple time scales were used to predict healthcare conditions. Using neural network predictive models which are based on a combination of the embedding of events, Esteban[11] predicted the clinical events recorded in the electronic medical record for the one who suffered from kidney failure.

In the medical area, Recurrent Neural Networks (RNNs) is mostly used in predicting temporal events. Similarly to our problem, Lipton [20] used LSTM to predict the diagnoses formulated by a multi-label classification problem. The better performance than the chosen baseline was achieved by replicating the classification target at each step. Regarding kidney failure, RNN and its variants were also used to predict the occurrence of diagnoses event within the next six or twelve months after the clinic visits [12]. DoctorAI proposed in [6] used longitudinal patient visit records to predict the physician diagnosis and medication order of the next visit. A 3-layer RNN was built to predict the diagnosis and medication categories for a subsequent visit. Recently, attention mechanism was introduced to improve the RNN's performance in some specific clinical problems. These models would achieve better performances because they focused on the critical information on a given input. Ma [21] introduced an attention-based bidirectional RNN using patient EMRs data to learn the relationship between subsequent visits. Choi [8] designed a two-level neural attention model with good clinical interpretability. One was for visit-level and the other for variable-level. Baytas [3] proposed a new LSTM unit with particular attention on the time intervals between two clinical events in longitudinal patient records.

6 CONCLUSION

In this paper, we have designed a novel healthcare service by developing an intelligent treatment engine, which can provide a patient with next-period prescriptions automatically and individually in real time. In order to cope with the complexity of medical practice and EMR data due to sequences having different lengths and record frequencies with multiple types of inter-correlations, a new LSTM learning framework has been proposed, so as to construct sequential hidden states for each medical sequences, model their connections with links between hidden neurons, and incorporate static factors with dynamic hidden states. In doing so, three multi-faceted LSTM models have been developed with fully connected, directional, and decomposed internal connections, respectively. Finally, experimental results validated the effectiveness of the proposed models. Particularly, the decomposed heterogeneous LSTM achieved the highest ROC-AUC on all 12 datasets and the highest PR-AUC on 11 out of the 12 datasets. Furthermore, future research will be undertaking from both theoretical and practical perspectives. Theoretically, we will explore how to further improve the performance of the heterogeneous LSTM models by introducing ontology of medicines or knowledge graph of laboratory indicators. In practice, we will apply the treatment engine in real-world applications and compare its results with that of doctors.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (No.61772110, 91546123, 71490724), China Postdoctoral Science Foundation (No. 2017M620054), and the National Key R&D Plan by the Ministry of Science and Technology of China (No.2016YFC0901900).

REFERENCES

[1] Arvind Balasubramanian, D Kim, Y Cheung, A Sawant, and B Prabhakaran. 2014. Analysis of Surface Motion Patterns Changes for Detecting Baseline Shifts in Respiratory Tumor Motion Data. In *3rd Workshop on Data Mining for Medicine and Healthcare (DMMH), 14th SIAM International Conference on Data Mining (SDM 2014), Philadelphia, USA.*

[2] The World Bank. 2016. Health Expenditure of the United States. http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS?locations=US. (2016).

[3] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 65–74.

[4] Vikas Chaurasia and Saurabh Pal. 2017. Data mining techniques: To predict and resolve breast cancer survivability. (2017).

[5] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 507–516.

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

[7] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1495–1504.

[8] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems.* 3504–3512.

[9] Clayton Christensen, Jerome H. Grossman, and M.D. Hwang. 2008. The Innovator's Prescription: A Disruptive Solution to the Healthcare Crisis. *McGraw-Hill* (2008).

[10] Saba Emrani, Anya McGuirk, and Wei Xiao. 2017. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1457–1466.

[11] Cristóbal Esteban, Danilo Schmidt, Denis Krompaß, and Volker Tresp. 2015. Predicting sequences of clinical events by using a personalized temporal latent embedding model. In *Healthcare Informatics (ICHI), 2015 International Conference on.* IEEE, 130–139.

[12] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on.* IEEE, 93–101.

[13] Rui Henriques, S Pina, and Cláudia Antunes. 2013. Temporal mining of integrated healthcare data: Methods, revealings and implications. *SDM IW on data mining for medicine and healthcare* (2013), 52–60.

[14] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. (2001).

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

[16] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. 2017. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing.* 17–33.

[17] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3 (2016).

[18] Wonsung Lee, Youngmin Lee, Heeyoung Kim, and Il-Chul Moon. 2016. Bayesian Nonparametric Collaborative Topic Poisson Factorization for Electronic Health Records-Based Phenotyping.. In *IJCAI.* 2544–2552.

[19] Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. 2017. NEMO: Next Career Move Prediction with Contextual Embedding. In *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, 505–513.

[20] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. *Computer Science* (2015).

[21] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1903–1911.

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

[23] Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek V Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed Memory Networks for Clinical Diagnostic Inferencing.. In *AAAI.* 3274–3280.

[24] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Peter J Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, et al. 2018. Scalable and accurate deep learning for electronic health records. *arXiv preprint arXiv:1801.07860* (2018).

[25] James C Ross, Peter J Castaldi, Michael H Cho, Junxiang Chen, Yale Chang, Jennifer G Dy, Edwin K Silverman, George R Washko, and Raúl San José Estépar. 2017. A Bayesian Nonparametric Model for Disease Subtyping: Application to Emphysema Phenotypes. *IEEE transactions on medical imaging* 36, 1 (2017), 343–354.

[26] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, NY, USA, 1865–1874.

[27] Ieva Vasiljeva and Ognjen Arandelovic. 2016. Automatic knowledge extraction from EHRs. In *IJCAI 2016-Workshop on Knowledge Discovery in Healthcare Data.*

[28] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining Electronic Health Records (EHRs): A Survey. *ACM Computing Surveys (CSUR)* 50, 6 (2018), 85.

[29] Sen Yang, Xin Dong, Leilei Sun, Yichen Zhou, Richard A. Farneth, Hui Xiong, Randall S. Burd, and Ivan Marsic. 2017. A Data-driven Process Recommender Framework. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, NY, USA, 2111–2120.

[30] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* ACM, 729–732.

[31] Zhongtang Zhao, Yiqiang Chen, Junfa Liu, Zhiqi Shen, and Mingjie Liu. 2011. Cross-people mobile-phone based activity recognition. In *IJCAI*, Vol. 11. 2545–250.