J. Vis. Commun. Image R. 72 (2020) 102901

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Real-time sepsis severity prediction on knowledge graph deep learning networks for the intensive care unit☆

Qing Li [a], Lili Li [b], Jiang Zhong [c], L. Frank Huang [a,*]

[a] Division of Experimental Hematology and Cancer Biology, Brain Tumor Center, Cincinnati Children's Hospital Medical Center, Cincinnati, United States
[b] School of Civil Engineering, Chongqing University, Chongqing, China
[c] College of Computer Science, Chongqing University, Chongqing, China

ARTICLE INFO

Keywords:
Deep neural networks
Sepsis
Intensive care units
Clinical informatics
Illness severity prediction
Knowledge graph

ABSTRACT

Sepsis is the third-highest mortality disease in intensive care units (ICUs). In this paper, we proposed a deep learning model for predicting the severity of sepsis patients. Most existing models based on attention mechanisms do not fully utilize knowledge graph based information for different organ systems, such that might constitute crucial features for predicting the severity of sepsis patients. Therefore, we have employed a medical knowledge graph as a reliable and robust source of side information. End-to-end neural networks that incorporate analyses of various organ systems simultaneously and intuitively were developed in the proposed model to reflect upon the condition of patients in a timely fashion. We have developed a pre-training technique in the proposed model to combine it with labeled data by multi-task learning. Experimental results on real-world clinical datasets, MIMIC-III and eIR, demonstrate that our model outperforms state-of-the-art models in predicting the severity of sepsis patients.

1. Introduction

Sepsis is a very significant disease in the intensive care units (ICUs), not to mention how burdensome it is financially. It accounts for more than $6 billion pounds in hospital expenses of UK [1]. In 2011 the US spent over 20 billion dollars on hospital care for sepsis patients, with expense growing to over 23 billion (responsible for 6.2% of all US hospital costs) [2]. Sepsis is defined as a severe "life-threatening organ dysfunction caused by a dysregulated host response to infection" [3–5]. Other than general guidelines (the Surviving Sepsis campaign), no tool is currently available for personalizing the treatment of sepsis [6–8]. The clinicians' task of then of deciding on the treatment type and dosage for individual patients is highly challenging. It is even more difficult to make decisions from the patient perspective. Moreover, the clinical variability in sepsis treatment is extreme, with clear evidence that suboptimal decisions cause poor outcomes.

In the ICU, the availability of more than 18.8 million electronic health records (EHRs) data presents a unique opportunity for generating novel insights in the effort to figure out how to facilitate better care [9]. Simultaneously, the penetration of EHRs into our procedures has seen an explosion in growth in the United States, from 9.4% (in 2008) to 83.8% (in 2015), a 9-fold increase [10], and thereby exceeding expectations. Efficient use of EHRs data for translational research is becoming increasingly significant, as it can help to make more precise, robust, and personalized decisions. But to date, minimal research has focused on predicting the severity of sepsis in patients based upon available EHRs data.

Recently, many meaningful machine-learning studies have been conducted on images and natural language processing [11–13]. These studies proposed many end-to-end artificial intelligence models that have enabled computer scientists to develop innovative decision support systems for the ICU. The current analytics on EHRs data of sepsis patients are focused mainly on mortality risk prediction [14,15] and on the use of antibiotics that tends to improve survival chances [16]. Neither means can support important physiological decision-making insights on dynamic change for clinicians. Severity score systems, such as SAPS (simplified acute physiology score) [17], SAPS-II [18], APS-III (acute physiology score) [19], OASIS (oxford acute severity of illness score) [20], APACHE-II (acute physiology and chronic health evaluation score) [21], and SOFA (sequential organ failure assessment score) [22] have been developed with the objective of predicting hospital mortality based on baseline patient characteristics. However, evaluations of scoring systems take a long time, and the results may be measured within 24 h after ICU admission. It is the heartfelt preference of patients and their families, and their families though to be provided

☆ This paper has been recommended for acceptance by Prof Xu Xing.

* Corresponding author.Tel.: +1-513-517-1084; fax: +1-513-803-5490;
E-mail address: Frank.Huang@cchmc.org (L. Frank Huang).

https://doi.org/10.1016/j.jvcir.2020.102901
Received 12 July 2019; Received in revised form 16 June 2020; Accepted 3 September 2020
Available online 6 September 2020

Table 1
Organ dysfunction in sepsis.

Important target organ in the system	Pathophysiology	Clinical features	SOFA score indices = 0
Respiratory system — Lung (ARDS)	Vascular hyper-permeability, neutrophil accumulation	Impaired oxygenation	$PaO_2/FIO_2 \geqslant$ 400 mmHg
Nervous system (SAE)	Direct cellular damage, mitochondrial and endothelial dysfunction, neurotransmission disturbances, calcium dyshomeostasis	Altered mental status	GCS=15
Cardiovascular system	Myocardial depression, impaired intracellular calcium homeostasis, disrupted high energy phosphate production.	Ventricular dilatation, reduced ejection fraction, reduced contractility	Mean arterial pressure ⩾70 mmHg
Blood coagulation system (DIC)	Intravascular coagulation, microvascular damage, systemic thrombin generation, endothelial injury	Bleeding diathesis, microthrombi and tissue ischemia	Platelets $\geqslant 150\times10^3$/μl
Digestive system — Liver	Disturbed intracellular and extracellular bile salt transport	Jaundice, cholestasis	Serum bilirubin < 20 [μmol/L]
Urinary system — Kidney (AKI)	Tubular epithelial cell injury, dysfunction or adaptive response of tubular epithelial cells	Reduced glomerular filtration ratio, reduced urine volume	Serum creatinine < 110 [μmol/L]

SOFA sequential organ failure assessment, ARDS acute respiratory distress syndrome, SAE sepsis-associated encephalopathy, GCS Glasgow coma scale, DIC disseminated intravascular coagulation, AKI acute kidney injury.

real-time feedback. Real-time feedback can calm these families down and help them to understand the ramifications of the condition rationally. At the same time, a critically ill patient cannot always wait much longer for the results of scoring system. The more frequent the routine evaluations, the more clearly the medical condition can be visualized, thereby providing a potential pathway for the early prediction of a medical condition.

Over the past few years, much of the existing literature on deep models has more densely predicted the illness severity score. The researchers focus on the most important multivariate time-series information in the data, and they obtain state-of-the-art results in a many severity prediction tasks. However, they focus only on physiological time-series features, such as the overall input from the human body, while ignoring the correlation between separate organ systems. Table 1 summarizes the pathophysiology, clinical features, and SOFA score indices. It can be seen that when SOFA = 0, the correlation between separate organ systems is associated with poor outcomes. It may negatively impact the prediction performance. This may be because these studies are insensitive to abnormal physiological characteristics if related organs begin to fail. But at this time, their failure will more easily lead to simultaneous failure of other organs.

In order to address these issues, we have proposed a data-driven approach that uses analyses of various organ systems to identify or come up with optimal sepsis treatment strategies. We combined pre-training bidirectional LSTM networks with the knowledge graph to predict severity of sepsis patients and how best to treat them in the ICU to improve their likelihood of survival. Our main contributions are summarized as follows:

- Based on a prior pre-training technique namely, MT-DNN [23], the model combines the temporal correlations between organ systems via multi-task learning. This model helps to collect complex temporal correlations data benefit from what we have collected related organ systems captured in MT-DNN;
- Furthermore, we proposed an end-to-end recurrent neural model (RNN) and attention mechanism with each learning unit, to focus on important input information and latent relations. Via this approach, we took advantage of sub-tasks that have fine-grained analysis of the data relations;
- Moreover, the model links medical knowledge maps for the first time. The model simulates the propagation of separate organ systems over a presented set of knowledge entities by automatically propagating potential influences of sepsis progression along with links in the knowledge graph. We present novel benchmark results by comparing many state-of-the-art methods of deep learning models on the MIMIC-III and eIR datasets.

The remainder of this paper is organized as follows: In Section 2, we discuss the related work. In Section 3, we introduce and compare types of datasets. In Section 4, we describe our method in detail. Comprehensive experiments and analyzes are presented in Section 5. Finally, we discuss the conclusions of this study.

2. Related work

Large-scale EHRs sepsis identification mainly relies on time-series data that store various types of medical variables. In intensive care units, sepsis appears to present usable and viable criteria for retrospectively realizing real-time severity prediction in EHRs. The main reasons are as follows: (1) it is timely, (2) it is consistent with other criteria, and (3) it satisfies many forms of validity [24]. Time-series models have been used in EHRs data analysis for various learning tasks, such as patient mortality prediction, disease progression, medical risk evaluation, and disease diagnosis modeling.

Over the past few years, early methods used handcrafted features via a series of analysis tools to design many sophisticated models. However, Caruana et al. [25] and Cooper et al. [26] found that AI models obtain better results on patient mortality prediction and medical risk evaluation. In order to improve prediction performance of AI model, Ghassemig et al. [27], Pirracchio et al. [28], Johnson et al. [29], and Wang et al. [30] exploit the correlations that are embedded in high-level features to predict the disease progression and phenotype of ICU patients. Unfortunately, due to the discreteness of time-series data and the underutilization of temporal information, the performances of these methods are limited. Physionet,[1] a competition platform, invites participants to develop machine learning models for addressing open health care problems.

With the recent advances in deep learning techniques, researchers have exhibited a growing interest in applying these techniques to healthcare applications due to the increasing availability of large-scale health care data [31–33]. Deep learning techniques significantly alleviates the tedious traditional work on feature engineering and extraction. In addition, researchers have found that temporal features that reflect changes in patients' conditions in a real-time setting can improve performance [34,35]. Che et al. [36] and Lipton et al. [37] modified the standard GRU and LSTM models to handle irregular EHRs time-series data for medical predict for patients. Weng et al. [38] uses time-series Q-learning to address glycemic control issues for sepsis patients. Hharutyunyan et al. [39] have developed a framework that is based on EHRs time-series RNNs for predicting in-hospital mortality. Aczon et al. [40] proposed pRNN, which is a system of pediatric encounter record (including physiologic observations, laboratory results, drugs,

[1] http://physionet.org/challenge/.

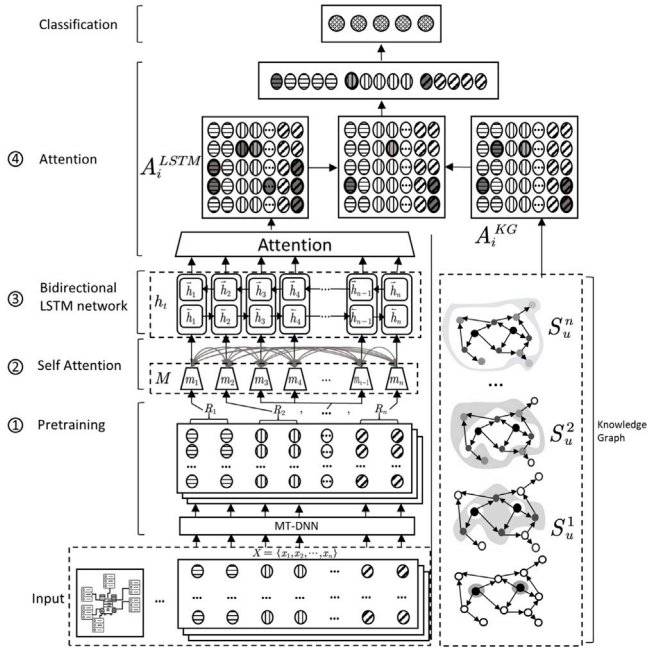

Fig. 1. Architecture of our model.

and interventions) with an RNN framework for prediction mortality. Chen et al. [41] proposed RNN-ATT, which is a recurrent neural network with an attention mechanism for predicting the illness severity scores.

These methods are ineffective because the "time" information uses monotonically decreasing functions heuristically. The above model focuses only on multi-task learning in clinical analytics to capture the relationships between medical tasks, which causes over-parameterization or under-parameterization in modeling time intervals. In addition, these methods fail to use various patterns of connections among separate organ systems from medical knowledge to provide additional guidance for recommendations. Currently, no model is available that can focus on the treatment of sepsis and assist clinicians in making real-time decisions at the patient level.

3. Methodology

3.1. Module overview

In this section, we introduce the proposed model framework, which incorporates a pre-training mechanism with a bidirectional LSTM network on a knowledge graph in detail. As illustrated in Fig. 1, our model consists of five main components: pre-training, self-attention, bidirectional LSTM network, knowledge graph, attention. Our model will not only learn the unique features from human organ systems in time-series EHRs and but also utilize the temporal correlations between organ systems. The most important characteristic of the model is that the conditions of ICU patients can be predicted in real-time using the SOFA score.

Table 2 provides an overview of the notation that we will use in this study.

3.2. Pre-training

Prior to applying the pre-training, we preprocessed the EHRs data of MIMIC-III, this process included cohort selection, data extraction, data cleaning, and feature extraction. The illustration of data preprocessing step are illustrated in Fig. 2. For each sepsis patients' medical record,

Table 2
Overview of main notations.

Notation	Definition
$\boldsymbol{x}_i$	A quarter hour of illness severity score in ICU stay record
R_i	Each organ system feature
$a_{t_i,d}$	Record of the t_i-th time ICU stay in the d-th organ system feature
h_t	The hidden state of Bi-LSTM
$\overrightarrow{h}_t$	The hidden state of forward LSTM network
$\overleftarrow{h}_t$	The hidden state of backward LSTM network
$\mathcal{G}$	The medical knowledge graph
(b, r, e)	Head entity – Relation – Tail entity
A_i^{KG}	The knowledge graph entity attention matrix
A_i^{LSTM}	The LSTM entity attention matrix

we extracted 48 features from human organ systems that are defined by MIMIC-III. All the extracted features would then be converted into a matrix with a variable number of rows. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ denote the input data, where n is the number of ICU records. We transformed each human organ systemic feature into vector representations by looking up embedding matrix $\boldsymbol{x}_i \in \mathbb{R}^{t_i \times D}$, where t_i is the maximum time length for the ith data sample ($i = 1, \ldots, n$) and D is the number of features.

According to Purushotham et al. [42] raw data were often of poor quality due to noise, missing values, and outliers, among other factors, and thereby we used the forward-fill strategy to optimize the raw data. For the missing value of $a_{t,d}$ ($d = 1, \ldots, D$), if all previously recorded missing values are still missing, they were then replaced by the median value of overall measurement. If there is at least one valid observation at time $t' < t$, replace $a_{t,d} := a_{t',d}$.

Instead of initializing the pre-training with BERT, we initialized the pre-training on the encoder of our model with MT-DNN [23]. It has the same architecture as BERT, but it is trained on multiple GLUE tasks. Herein, we extend MT-DNN to encode the data, and "[SEP]" is used to separate the data between ICU records, namely, each data input is encoded as: [CLS], $\boldsymbol{x}_1$, [SEP], $\boldsymbol{x}_2$, … [SEP], $\boldsymbol{x}_n$, [SEP]

[CLS] emphasizes that data information is being captured along time-series position, with [SEP] providing special token for data separation. We attempt to learn more time-series information by strengthening the original output of encoder pre-training with the global association information that is captured by [CLS].

3.3. Self-attention

Same systems data representations are differently fixed in time-series, even though the meanings of the data vary, depending on the human organ systems. Many neural network models that used to encode sequences of data may expect to learn implicit meanings of the human organ systems, however they cannot learn well due to long-term dependency problems. To overcome this problem, we employ representation vectors of self-attention to capture the meaning of time-series data that considering the human organ systems.

We used the multi-head attention formulation [43] for implementing self-attention. As illustrated in Fig. 3, self-attentions consists of several linear transformations and a scaled dot-product attention.

Given a matrix with n vectors, each organ system data R_i from beginning to end occurs at a different time $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ in ICU stay records, as sepsis patients may be hospitalized more than once in the ICU, other system training data *Other*. The scaled dot-product attention is calculated via the following equation:

$$Attention\left(R_{i_{begin}}, R_{i_{end}}, Other\right) = softmax\left(\frac{R_{i_{begin}} R_{i_{end}}^{\top}}{\sqrt{d_D}}\right) Other. \quad (1)$$

where R and *Other* are equivalent to X, where d_D is the dimension of the vector. The scaled dot-product attention with linear transformations is performed on r parallel heads to pay attention to different parts, defined as follows:

$$MultiHead(R_{i_{begin}}, R_{i_{end}}, Other) = W^D \left[head_1; \ldots; head_r\right], \quad (2)$$

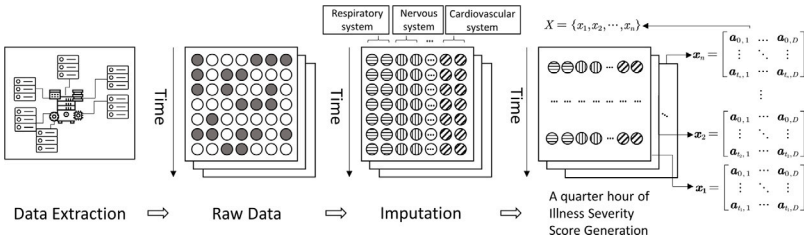

Fig. 2. Illustration of the input data structure and organization.

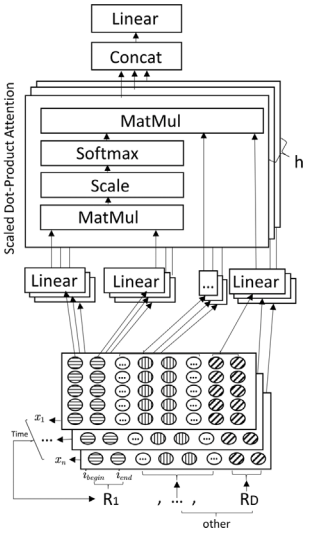

Fig. 3. Illustration of self-attention.

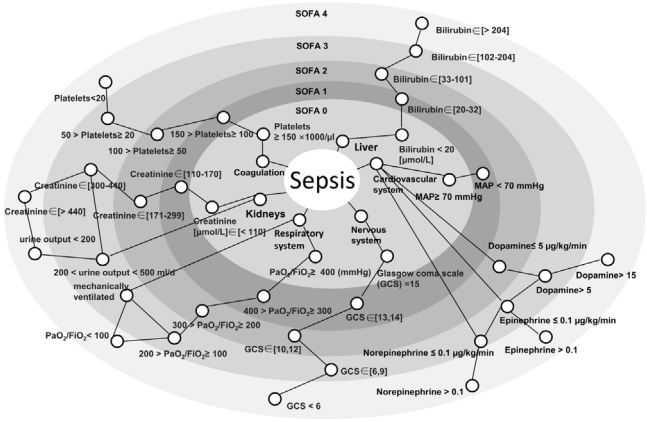

Fig. 4. Illustration of the medical knowledge graph mechanism.

where [;] represents row concatenation, r is the number of heads, and $W^D \in \mathbb{R}^{t_i \times D}$.

$$head_i = Attention\left(W_i^{R_{i_{begin}}} R_{i_{begin}}, W_i^{R_{i_{end}}} R_{i_{end}}, W_i^{Other} Other\right), \tag{3}$$

where $W_i^{R_{i_{begin}}} \in \mathbb{R}^{d_D/r \times d_D}$, $W_i^{R_{i_{end}}} \in \mathbb{R}^{d_D/r \times d_D}$, $W^{Other} \in \mathbb{R}^{d_D/r \times d_D}$. The outputs are for concatenation of the scaled dot-product attention. Hence, the outputs of multi-head attention are denoted by $D = \{m_1, m_2, \ldots, m_n\} = MultiHead(R_1, \ldots, R_D)$.

3.4. Bidirectional LSTM network

We use a Bi-LSTM network [44,45] that consists of forward and backward sub-LSTM networks:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right], \tag{4}$$

$$\overrightarrow{h_t} = \overrightarrow{LSTM}\left(m_t\right), \tag{5}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}\left(m_t\right), \tag{6}$$

where h_t is the hidden state for time step t, which it the concatenation of the $\overrightarrow{h_t}$ (forward) and $\overleftarrow{h_t}$ (backward) sub-LSTM networks ($h_t \in \mathbb{R}^{2d_h}$, $\overrightarrow{h_t} \in \mathbb{R}^{d_h}$, and $\overleftarrow{h_t} \in \mathbb{R}^{d_h}$). The representation vectors of $D = \{m_1, m_2, \ldots, m_n\}$ that are obtained from the self-attention layer are forwarded into the network step by step.

3.5. Knowledge graph

As illustrated in Fig. 4, a medical knowledge graph typically contains fruitful facts and connections among separate organ systems.

In a knowledge graph $\mathcal{G}$, let (b, r, e) denote the set of massive entity(head)-relation-entity(tail) triples. Herein $b \in \mathcal{E}$, $r \in \mathcal{R}$, and $e \in \mathcal{E}$ denote the head, relation, and tail of a knowledge triple, respectively, $\mathcal{E}$ and $\mathcal{R}$ denote the set of entities and relations in the knowledge graph (KG). For the input entity v, the set of i relevant entities for the user is defined as:

$$S_u^i = \left\{(b, r, e) \mid (b, r, e) \in \mathcal{G} \text{ and } b \in \mathcal{E}_u^{i-1}\right\}, i = 1, 2, \ldots, n. \tag{7}$$

Given the input entity embedding v and the 1-hop ripple set S_u^1, each triple (b_i, r_i, e_i) in S_u^1 is assigned a relevance probability p_i by comparing item v to the head b_i and the relation r_i in the triple:

$$p_i = Softmax\left(\mathbf{v}^T \mathbf{R}_i \mathbf{b}_i\right) = \frac{\exp\left(\mathbf{v}^T \mathbf{R}_i \mathbf{b}_i\right)}{\sum_{(b,r,e) \in S_u^1} \exp\left(\mathbf{v}^T \mathbf{R} \mathbf{b}\right)}, \tag{8}$$

where $\mathbf{R}_i \in \mathbb{R}^{d \times d}$ denotes the embedded matrix of relations r_i, and $\mathbf{b}_i \in \mathbb{R}^d$ denotes the embedded matrix of b_i.

By combining the responses of all ripple sets, the matrix of knowledge graph entity attention is obtained:

$$A_i^{KG} = \sum_{i=1}^{n} \sum_{(h_i, r_i, t_i) \in S_u^1} p_i \mathbf{e}_i. \tag{9}$$

3.6. Attention

Many attention models have been widely used in computer vision and natural language tasks to "focus on" a region of information while forming perceiving of what surrounding information is presented, and the focal point is adjusted over time. As illustrated in Fig. 1, we proposed an attention mechanism for learning the important features selectively, and we established direct short-cut connections between the target and the source. For each patient, we calculated the attention weight using the dot-product of the hidden state for every feature in the input, where the ith feature is computed as follows:

$$u_i = h_f^{\top} \hat{h}_s, \tag{10}$$

Table 3
Description of the datasets.

	MIMIC-II	MIMIC-III	eRI
Unique ICU admissions (n)	15,268	17,083	79,073
Characteristics of hospitals, per number of ICU admissions	Teaching tertiary hospital	Teaching tertiary hospital	Nonteaching: 37,146 (47.0%) Teaching: 29,388 (37.2%) Unknown: 12,539 (15.9%)
Age, years (mean)	64.6	64.4	65
Male gender (n (%))	8,175(53.5%)	9,604 (56.2%)	40,949 (51.8%)
Primary ICD-9 Sepsis diagnosis (n (%))	5,185(33.9%)	5,824 (34.1%)	41,396 (52.3%)
Initial OASIS (mean)	32.9	33.5	34.8
Initial SOFA (mean)	6.34	7.2	6.4
Length of stay, days (median)	2.9 (1.7–7)	3.1 (1.8–7)	2.9 (1.7–5.6)
ICU mortality	6.88%	7.40%	9.80%
Hospital mortality	9.85%	11.30%	16.40%
90-d mortality	17.82%	18.90%	–

where $h_f^{\top}$ is the learned feature of the input, and $\hat{h}_s$ is the concatenated hidden state, in which the i-th feature of the input is imputed.

$$W_i = \frac{\exp\left(u_i\right)}{\sum_{j=1}^{n} \exp\left(u_j\right)}, \tag{11}$$

$$A_i^{LSTM} = \sum_{t=1}^{n} W_i h_i, \tag{12}$$

$$A_f = A_i^{LSTM} + A_i^{KG}, \tag{13}$$

where W_i denotes the weight for the t-th feature, j contains all input features, and A_f denotes the final state output.

4. Experiments

4.1. Dataset

In this subsection, we describe the characters of MIMIC-II, MIMIC-III, and eRI datasets. Additional details and comparisons are provided in Table 3.

- **MIMIC-II:** The MIMIC-II dataset [46] includes physiologic information for all critical care admissions between 2001–2007 at BIDMC in Boston, and it is maintained by researchers at the Harvard-MIT (Massachusetts Institute of Technology) Division of Health Sciences and Technology.[2] ICU nurses have validated these collections of data on an hourly basis, such as regarding measurements of blood pressure, heart rate, IV medications, and also nursing progress notes, fluid intake/output, provider order entry (POE), ICD-9 codes, and ECG reports.
- **MIMIC-III:** The MIMIC-III dataset [47] is an extension of MIMIC-II, form which all 48 features were extracted, and it contains admissions from an additional four years (2008–2012). This dataset encompasses a diverse and very large population of ICU patients. Data were included from up to 24 h preceding the diagnosis of sepsis and up to 48 h following the onset of sepsis. This represents derived from early phase of initial resuscitation. This dataset is freely available to researchers worldwide.[3]
- **eRI:** The database of the eICU Research Institute (eRI),[4] which is a nonprofit institute that was established by Philips and is governed by customers. This is a platform built from a repository of data used to advance knowledge of critical and acute care. It contains more than 3.3 million admissions from 2003–2016 in 459 ICUs across the United States.

[2] http://physionet.org/mimic2.
[3] https://mimic.physionet.org/about/mimic/.
[4] https://www.usa.philips.com/healthcare/solutions/enterprise-telehealth/eri.

4.2. Experimental design and dataset

We evaluated our model on the MIMIC-III and eRI datasets, which are commonly-used benchmark datasets, and we compared the results obtained by the proposed model with state-of-the-art models. More comparison results are presented in Table 3. In both MIMIC-III and eRI, we trained and tested our model on data for adult patients that satisfied the international consensus sepsis-3 criteria [3]. Our model implementation was built by using the MT-DNN, which extends the BERT implementation in PyTorch. The remaining weights were randomly initialized from a zero-mean Gaussian distribution. We applied a cross-validation procedure to tune the hyper-parameters of our model on the development set. After data cleaning, we got 17,083 admissions of eligible patients with sepsis from MIMIC-III and 79,073 admissions from eRI. The dataset contains 48 distinguished features, which are listed in Table 4.

4.3. Experimental results on the MIMIC-III dataset

First, we randomly selected 80% of the MIMIC-III dataset and have trained the methods on it, while the remaining 20% of the dataset was used as the testing sets. Second, we have trained all the methods on the MIMIC-III dataset and tested on the independent dataset taken from the eRI database. These datasets have up to 72 h of measurements from approximately estimated time of the onset of sepsis. The data was then coded as multidimensional discrete-time series with 4 h time steps. The loss function was entropy-based function and used an L2-norm regularization term.

Table 5 compares the characteristics of existing benchmarking models. We have divided these models into three groups: traditional model, neural model, and our model. Table 6 compares our pre-training bidirectional LSTM networks model with the state-of-the-art models that have been applied to the MIMIC-III dataset.

First, the traditional models used many handcrafted features for classification, such as SVM(support vector machine), RF(random forest), DT(decision tree), LDA(linear discriminant analysis), Bayesian classification approaches [48,49] and XGboost [50]. As a result, they realized an F1-score of 72.30%. The neural model is capable of catching more features for detecting data relationships within different organ systems. The neural models may not however always be accurate and the parsing time is exponentially increased by the incoming data size. RNN-based models, such as GRU [36], RNN [39], PRNN [40], and RNN+ATT [41], are available for this task. We have used the RNN+ATT model for pre-training bidirectional LSTM networks and learned internal representations that occur between the original inputs and the final outputs in deep learning. We tuned all the hyper-parameters for our model via 5-fold cross-validation to the optimize of the hyper-parameters. The experimental results demonstrate that our model achieved an improvement of 3.25% in the F1-score over the second-best RNN+ATT model.

The best hyper-parameters are presented in Table 7.

Table 4
List of 48 features in the datasets.

Category	Items	MIMIC-III	eRI
Demographics	Age	Y	Y
	Gender	Y	Y
	Weight	Y	Y
	Readmission to intensive care	Y	Y
	Elixhauser score (premorbid status)	Y	**N**
Vital signs	SOFA (4h time step)	Y	Y
	SIRS	Y	Y
	Glasgow coma scale	Y	Y
	Heart rate	Y	Y
	systolic	Y	Y
	mean and diastolic	Y	Y
	blood pressure	Y	Y
	shock index	Y	Y
	Respiratory rate	Y	Y
	SpO_2 Temperature	Y	Y
Lab values	Potassium	Y	Y
	sodium	Y	Y
	chloride	Y	Y
	Glucose	Y	Y
	BUN	Y	Y
	creatinine	Y	Y
	Magnesium	Y	Y
	calcium	Y	Y
	ionized calcium	Y	Y
	carbon dioxide	Y	Y
	SGOT	Y	Y
	SGPT	Y	Y
	SGPT	Y	Y
	albumin Hemoglobin	Y	Y
	White blood cells count	Y	Y
	platelets count	Y	Y
	Partial Thromboplastin Time	Y	Y
	Prothrombin Time	Y	Y
	International Normalized Ratio	Y	Y
	pH	Y	Y
	PaO_2	Y	Y
	$PaCO_2$	Y	Y
	base excess	Y	Y
	bicarbonate	Y	Y
	lactate	Y	Y
Ventilation parameters	Mechanical ventilation	Y	Y
	FiO_2	Y	Y
Medications and fluid balance	Current IV fluid intake over 4h	Y	Y
	Maximum dose of vasopressor over 4h	Y	Y
	Urine output over 4h	Y	Y
	Cumulated fluid balance since admission	Y	Y
Outcome	Hospital mortality	Y	Y
	90-day mortality	Y	**N**

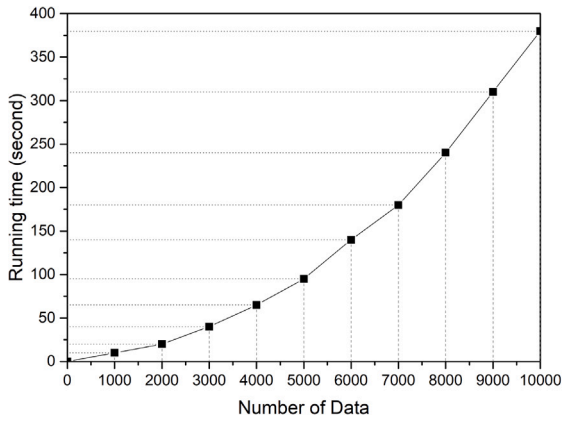

Fig. 5. Averaged running time records with the increase in the number of patients data.

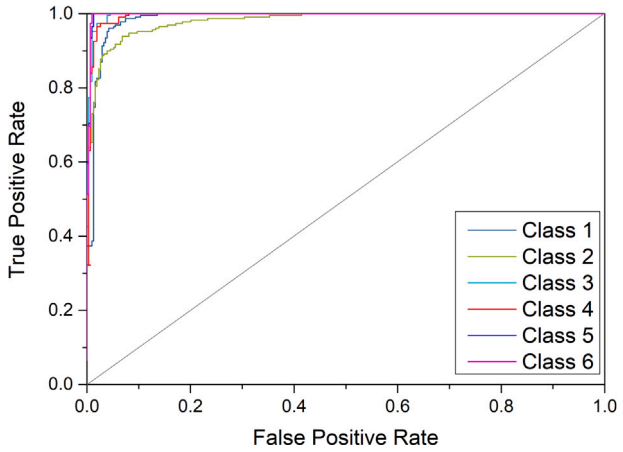

Fig. 6. The ROC curves of different classifications predicated by our model on MIMIC-III dataset.

To investigate the importance of each component of our model, we conducted ablation experiments. From the results, we observed that the pre-training mechanism to prepare bidirectional LSTM significantly outperforms RNN. At the same time, the pre-training module plays a highly important role in the experiment. As a result, the experimental results demonstrate that our model realized an F1-score of 93.15%; thus, it outperforms the state-of-the-art approaches. The pre-training module improves the performance by 1.49% compared to the model without its application. Similarly, the self-attention mechanism of the modules improves the performance by 0.87% compared with given models that have applied it. In the end, the attention mechanism of modules improves the performance by 1.62% in terms of F1-score. Based on these results, we found that the pre-training module makes the largest contributes most to the proposed model towards obtaining better F1-score. Meanwhile, the self-attention mechanism and attention mechanism may contribute to the exploitation of the temporal correlations between human organ systems by the shared hidden layer. The attention module has successfully extracted information from the medical knowledge graph in an effort to build a giant heterogeneous graphs regarding necessary facts in terms of diseases and medications for them. The proposed model used ICD-9 ontology to represents a knowledge base of human diseases, such that can be used to improve diagnosis and classify patient's disease. Harnessing well-built knowledge graphs might enable the reinvented system to provide suitable prescriptions for special patients, along with alerts regarding possible side effects and severe drug-drug interactions (DDIs). We have repeated the test 10 times and reported the averaged runtime records with an applied increase in the number of patient data, as shown in Fig. 5, where X-axis is the number of data and Y-axis represents the corresponding running time of our algorithm in seconds.

The SOFA scoring system is useful for predicting the clinical outcomes of critically ill patients with sepsis. We divided the severity of sepsis patients into six classes according to their SOFA score and mortality rate, as presented in Table 8.

Fig. 6 shows that the proposed model can accurate predict the severity of sepsis patients in terms of ROC curve. It also indicates that our method can effectively distinguish various types of intermediate conditions and predict various extreme scenarios of sepsis in ICU.

4.4. Experimental results on the eRI dataset

To evaluate the value of the proposed model, we applied the pre-training bidirectional LSTM networks to model the eRI database. As eRI dataset has much more samples than the MIMIC-III dataset, we can generalize the proposed more easily. Fig. 7 shows our model accurate

Table 5
The Comparison of benchmarking works.

Models	Cite	Time durations		Number of features		Databases				Scoring systems	
		24 h	48 h	Smaller	Larger	MIMIC-II	PICU	MIMIC-III	eRI	SAPS-II	SOFA
Super learner	(Pirracchio, 2016) [28]	Y		Y		Y				Y	Y
Reproducibility	(Johnson et al. 2017) [29]	Y	Y	Y				Y			
GRU	(Che et al., 2018) [36]		Y		Y			Y			
RNN	(Harutyunyan et al., 2017) [39]		Y	Y				Y			
pRNN	(Aczon et al., 2017) [40]	Y		Y			Y				
RNN-ATT	(Chen et al., 2018) [41]	Y			Y			Y			Y
Our model	(Li et al., 2019)	Y	Y		Y			Y	Y	Y	Y

Table 6
Comparison with the previous results on MIMIC-III dataset.

	Models	F1
Traditional model [50]	Support vector machine	68.93
	Random forest	71.53
	Decision tree	72.30
	Linear discriminant analysis	71.22
	XGboost	63.34
Neural model	GRU [36]	83.05
	RNN [39]	86.90
	PRNN [40]	80.41
	RNN+ATT [41]	89.90
Our model		**93.15**
- pretaining		91.66
- pretaining - Self-Attention		90.79
- pretaining - Self-Attention–Attention		89.17

Table 7
The best hyperparameters in our model.

Hyperparameters	Description	Value
d_D	Size of embedding	300
r	Number of heads	4
d_h	Size of hidden layer	300
λ	L2 regularization coefficient	10^{-5}
Dropout rate	Self attention layer	0.3
	Bi-LSTM layer	0.3
	Knowledge graph layer	0.3
	Attention layer	0.5

Table 8
Classification of the SOFA score.

Class	1	2	3	4	5	6
Sofa score	0–6	7–9	10–12	13–14	15	15-24
Mortality rate	10%	15%–20%	40%–50%	50%–60%	80%	90%

Table 9
Comparison with previous results on eRI dataset.

Models	F1
GLM [51]	82.02
XGboost [52]	83.19
RNN [53]	85.41
Our model	**88.32**
- pretaining	86.47
- pretaining - Self-Attention	84.95
- pretaining - Self-Attention–Attention	82.07

predicted the severity of sepsis patients in terms of ROC curve on the eRI dataset. We also find our method are with high sensitivity in predicting the sepsis patients that are assigned in "Class 1" and "Class 6", which are two extreme conditions in the ICU. Although the F1-score of the proposed model has decreased when the number of sepsis patients increased, the lowest F1-score achieved by our method is still larger than 85%.

Table 9 presents the F1-scores that are generated for early prediction using the GLM [51], XGBoost [52], and RNN [53] models. Our model realized an overall F1-score of 88.32%; thus, it outperforms all competing state-of-the-art approaches.

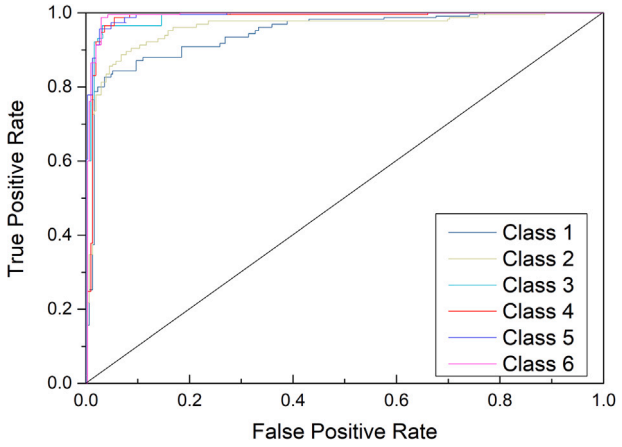

Fig. 7. The ROC curves of different classifications as illustrated by our model on eRI dataset.

Based on these results, we made the following observations:

- When using traditional features, the proposed model will yield higher performance in datasets with more feature. However, the quality of traditional features depends on the human originality and prior knowledge of the doctor, which are difficult to popularize. The pre-training model can solve this problem well. We conduct an ablation study to investigate how effectively it can place into context information — from the ensemble models to the pre-training. The experimental results demonstrate that our inference is correct.
- The reinforcement mechanism of attention modules can extract more abstract and higher-level features from the matrix. The self-attention mechanism and attention mechanism are very helpful for relational classification. Via this approach, relational influential data can be better located. At the same time, we conduct an ablation study to investigate this. The F1-score has also been improved this way.
- In experiments, we can always identify metaphorical descriptions in model training. With more data samples, we typically can train better models and come up with the best metaphorical descriptions. In addition, evidence-based medicine domain knowledge of sepsis can effectively facilitate the discovery of this association. We observed that this may lead to substantial improvements over these baselines.

5. Conclusions

In this paper, we proposed pre-training bidirectional LSTM networks with pre-training to predict the severity of sepsis patients in the ICU. Our model achieved an F1-score of 93.15% on the MIMIC-III dataset, outperforms the state-of-the-art methods. In addition, we have also

applied the proposed model to the eRI dataset. However, we found that the F1-scores were decreased on the eRI dataset. In the future, we will focus on developing new methods that use better medical knowledge graph to improve the prediction performance in terms of F1-score.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable comments.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or nonprofit sectors.

References

[1] J.-L. Vincent, Y. Sakr, C.L. Sprung, V.M. Ranieri, K. Reinhart, H. Gerlach, R. Moreno, J. Carlet, J.-R. Le Gall, D. Payen, Sepsis in european intensive care units: Results of the SOAP study, Crit. Care Med. 34 (2) (2006) 344–353.
[2] C. Torio, R. Andrews, National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160, Agency for Healthcare Research and Quality (US), Rockville (MD), 2006.
[3] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.-D. Chiche, C.M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), JAMA 315 (8) (2016) 801–810.
[4] V. Liu, G.J. Escobar, J.D. Greene, J. Soule, A. Whippy, D.C. Angus, T.J. Iwashyna, Hospital deaths in patients with sepsis from 2 independent cohorts, JAMA 312 (1) (2014) 90–92.
[5] J.E. Gotts, M.A. Matthay, Sepsis: Pathophysiology and clinical management, BMJ 353 (2016) i1585.
[6] L. Byrne, F. Van Haren, Fluid resuscitation in human sepsis: Time to rewrite history? Ann. Intensive Care 7 (1) (2017) 4.
[7] P. Marik, The demise of early goal-directed therapy for severe sepsis and septic shock, Acta Anaesthesiol. Scand. 59 (5) (2015) 561–567.
[8] J. Waechter, A. Kumar, S.E. Lapinsky, J. Marshall, P. Dodek, Y. Arabi, J.E. Parrillo, R.P. Dellinger, A. Garland, Interaction between fluids and vasoactive agents on mortality in septic shock: A multicenter, observational study, Crit. Care Med. 42 (10) (2014) 2158–2168.
[9] F.S. Collins, H. Varmus, A new initiative on precision medicine, N. Engl. J. Med. 372 (9) (2015) 793–795.
[10] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel, Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015, ONC Data Brief 35 (2016) 1–9.
[11] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: Spatial-temporal attention mechanism for video captioning, IEEE Trans. Multimed. (2019).
[12] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, Q. Dai, Cross-modality bridging and knowledge transferring for image understanding, IEEE Trans. Multimed. (2019).
[13] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast uyghur text detector for complex background images, IEEE Trans. Multimed. 20 (12) (2018) 3389–3398.
[14] F. Shann, G. Pearson, A. Slater, K. Wilkinson, Paediatric index of mortality (PIM): A mortality prediction model for children in intensive care, Intensive Care Med. 23 (2) (1997) 201–207.
[15] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, et al., The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults, Chest 100 (6) (1991) 1619–1636.
[16] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, Crit. Care Med. 46 (4) (2018) 547–553.
[17] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, JAMA 270 (24) (1993) 2957–2963.
[18] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group, JAMA 276 (10) (1996) 802–810.
[19] W. Knaus, The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults, Chest 102 (1992) 1919–1920.
[20] A.E. Johnson, A.A. Kramer, G.D. Clifford, A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy, Crit. Care Med. 41 (7) (2013) 1711–1718.
[21] W.A. Knaus, J.E. Zimmerman, D.P. Wagner, E.A. Draper, D.E. Lawrence, APACHE-Acute physiology and chronic health evaluation: A physiologically based classification system, Crit. Care Med. 9 (8) (1981) 591–597.
[22] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, L. Thijs, The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, Intensive Care Med. 22 (7) (1996) 707–710.
[23] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, 2019, arXiv preprint arXiv:1901.11504.
[24] A.E. Johnson, J. Aboab, J.D. Raffa, T.J. Pollard, R.O. Deliberato, L.A. Celi, D.J. Stone, A comparative analysis of sepsis identification methods in an electronic database, Crit. Care Med. 46 (4) (2018) 494–499.
[25] R. Caruana, S. Baluja, T. Mitchell, Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation, in: Advances in Neural Information Processing Systems, 1996, pp. 959–965.
[26] G.F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B.G. Buchanan, R. Caruana, M.J. Fine, C. Glymour, G. Gordon, B.H. Hanusa, et al., An evaluation of machine-learning methods for predicting pneumonia mortality, Artif. Intell. Med. 9 (2) (1997) 107–138.
[27] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, P. Szolovits, Unfolding physiological state: Mortality modelling in intensive care units, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 75–84.
[28] R. Pirracchio, Mortality prediction in the ICU based on MIMIC-II results from the super ICU learner algorithm (SICULA) project, in: Secondary Analysis of Electronic Health Records, Springer, 2016, pp. 295–313.
[29] A.E. Johnson, T.J. Pollard, R.G. Mark, Reproducibility in critical care: A mortality prediction case study, in: Machine Learning for Healthcare Conference, 2017, pp. 361–376.
[30] S. Wang, X. Chang, X. Li, G. Long, L. Yao, Q.Z. Sheng, Diagnosis code assignment using sparsity-based disease correlation embedding, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3191–3202.
[31] A. Oellrich, N. Collier, T. Groza, D. Rebholz-Schuhmann, N. Shah, O. Bodenreider, M.R. Boland, I. Georgiev, H. Liu, K. Livingston, et al., The digital revolution in phenotyping, Brief. Bioinform. 17 (5) (2015) 819–830.
[32] F. Dabek, J.J. Caban, A neural network based model for predicting psychological conditions, in: International Conference on Brain Informatics and Health, Springer, 2015, pp. 252–261.
[33] N.Y. Hammerla, J. Fisher, P. Andras, L. Rochester, R. Walker, T. Plötz, PD disease state assessment in naturalistic environments using deep learning, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[34] M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, A.A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Med. 24 (11) (2018) 1716.
[35] S. Purushotham, W. Carvalho, T. Nilanon, Y. Liu, Variational recurrent adversarial deep domain adaptation, 2016.
[36] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085.
[37] Z.C. Lipton, D.C. Kale, R. Wetzel, Modeling missing data in clinical time series with rnns, Mach. Learn. Healthc. (2016).
[38] W.-H. Weng, M. Gao, Z. He, S. Yan, P. Szolovits, Representation and reinforcement learning for personalized glycemic control in septic patients, 2017, arXiv preprint arXiv:1712.00654.
[39] H. Harutyunyan, H. Khachatrian, D.C. Kale, G.V. Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, 2017, arXiv preprint arXiv:1703.07771.
[40] M. Aczon, D. Ledbetter, L. Ho, A. Gunny, A. Flynn, J. Williams, R. Wetzel, Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks, 2017, arXiv preprint arXiv:1701.06675.
[41] W. Chen, S. Wang, G. Long, L. Yao, Q.Z. Sheng, X. Li, Dynamic illness severity prediction via multi-task RNNs for intensive care unit, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 917–922.
[42] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmark of deep learning models on large healthcare mimic datasets, 2017, arXiv preprint arXiv:1710.08531.
[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[44] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6645–6649.
[45] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.
[46] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.

[47] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.

[48] Q. Liu, L.J. Muglia, L.F. Huang, Network as a biomarker: a novel network-based sparse bayesian machine for pathway-driven drug response prediction, Genes 10 (8) (2019) 602.

[49] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, S.T. Wong, Drugcomboranker: drug combination discovery based on target network analysis, Bioinformatics 30 (12) (2014) i228–i236.

[50] S. Wang, X. Li, L. Yao, Q.Z. Sheng, G. Long, et al., Learning multiple diagnosis codes for ICU patients with local disease correlation mining, ACM Trans. Knowl. Discov. Data 11 (3) (2017) 31.

[51] P. McCullagh, Generalized Linear Models, Routledge, 2019.

[52] A. Grover, J. Leskovec, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016.

[53] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.