Predicting ICU Mortality from Heterogeneous Clinical Events with Prior Medical Knowledge

Lujing Xiao[1,2,3], Chuanpan Zheng[1,2], Xiaoliang Fan[1,2,3,*], Yi Xie[2], Rongshan Yu[1,3]
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China
[2]Computer Science Department, Xiamen University, Xiamen, China
[3]Digital Fujian Institute of Healthcare & Biomedical Big Data Research, Xiamen University, Xiamen, China
{xiaolujing, zhengchuanpan}@stu.xmu.edu.cn, {fanxiaoliang, csyxie, rsyu}@xmu.edu.cn

Abstract

Mortality prediction in Intensive Care Unit (ICU) could assist clinicians to make diagnosis. Many deep learning models have been previously proposed to uncover the high order correlations among heterogeneous clinical events. However, they failed to address the importance of prior medical knowledge from experienced clinicians. In this paper, we propose a novel ICU mortality prediction method called ***P-BiLSTM***, which combines a bidirectional Long Short-Term Memory model with prior medical knowledge of clinicians. In P-BiLSTM, we first preprocess general descriptors and heterogeneous temporal events with a sophisticated data completion strategy. After that, we use a weighted block with prior medical knowledge to enhance learning and explainable abilities of deep neural networks. The performance of the proposed method is validated using a real-world dataset containing 37 types of temporal events from 4,000 ICU patients within 48-hour. Experimental results demonstrate that P-BiLSTM outperforms six baseline methods.

1 Introduction

Medical diagnosis is a data-intensive and knowledge-intensive scenario, which requires strong abilities of knowledge reserving, processing and judgment [He *et al.*, 2019; Ma *et al.*, 2018]. For instance, clinicians in surgical ICU would pay attention to indicators such as *Platelets*. Because when the number of *Platelets* sharply reduces, it indicates that patients may encounter a life threatening issue such as postoperative bleeding. In contrast with the common-sense knowledge, clinicians in emergency department might be cautious about *hyperoxia* (i.e., using excessive oxygen) among mechanically ventilated patients. Because *hyperoxia* will generate toxic molecular in the blood that could be particularly injurious [Page *et al*., 2018]. In short, the lesson learnt from aforementioned examples is, besides continuous trends reflected from heterogeneous clinical events, clinicians make diagnoses largely based on their medical knowledge and experience. For example, they take into account the certain ICU type of a patient and underlying considerations accordingly. In addition, these problems encountered by clinicians are far more complex in real world cases. Thus, it is urgent to combine their rich medical knowledge with multiple clinical variables to make a precise diagnosis for a specific patient.

Many early works have utilized machine learning methods [Bhattacharya *et al.*, 2017; Citi and Barbieri, 2012] to optimize the prediction model with ICU datasets. More recently, many works tended to focus on mining the high order correlations among heterogeneous clinical variables, with the superior learning ability of deep neural networks such as Convolutional Neural Network (CNN) [Suo *et al.*, 2017], Long Short-Term Memory (LSTM) [Nguyen *et al.*, 2017] and many others [Yang *et al.*, 2016; Krizhevsky *et al.*,2012; Chung *et al.*, 2018]. However, they failed to address the importance of prior medical knowledge from experienced clinicians. In addition, existing works often conduct a straightforward strategy to deal with the missing data issue from heterogeneous temporal events. This is problematic since one causal factor of missing data relies largely on medical procedures, such as measuring blood pressure hourly, while collecting urine every 8 hours. As a result, a simple data completion strategy will inevitably introduce noises that might interfere with the prediction model.

To address the aforementioned challenges, we propose a novel ICU mortality prediction method, named ***P-BiLSTM***, which combines a bidirectional LSTM (Long Short-Term Memory) model with prior medical knowledge of experienced clinicians. First, we choose available general descriptors and time-series variables as the input of each patient with a sophisticated data completion strategy. Second, we design a weighted block with prior medical knowledge to enhance learning and explainable abilities of deep neural network model. Finally, we train and evaluate our model on a real-world dataset containing 37 types of heterogeneous temporal events from 4,000 ICU patients within 48-hour. Experimental results demonstrate that our proposed P-BiLSTM outperforms six baseline methods, including CNN, GRU, LSTM, BiLSTM, BiGRU and BiLSTM-ST.

2 Related Works

Recently, there are plenty deep neural network models to solve the clinical endpoint prediction problem with their capacity of mining high order correlations among multiple clinical variables [Liu *et al.*, 2018]. Convolutional neural network (CNN) shows strong ability to capture local features to predicting multiple diseases [Suo *et al.*, 2017]. Later, to capture longtime characteristics of patients' records, recurrent neural networks (RNNs) and its variants [Chung *et al.*, 2014; Yang *et al.*, 2016; Graves *et al.*, 2005; Gupta *et al.*, 2018] are applied to predict patients' health status based on electronic health records (EHR). Furthermore, Long Short-Term Memory (LSTM) network [Lipton *et al.*,2015] and its variants [Nguyen *et al.*, 2017; Zhu et al., 2018] are used to classify diagnoses based on massive EHR in pediatric intensive care units [Johnson *et al.*, 2016]. However, these works failed to combine rich medical knowledge from experienced clinicians with heterogeneous clinical events to make diagnoses and treatments for a specific patient. Instead, we propose a novel ICU mortality prediction method that not only enhances the predictive performance, but also makes the prediction result more explainable.

3 Preliminary

In this section, we describe the notation used in this paper, and the problem definition about a mortality prediction task.

3.1 Notation

Each patient p is associated with a specific *ICUType*, a sequence of heterogeneous events and *survival days*. *ICUType* is a categorical variable that specifies the type of ICU to which the patient has been admitted. We denote it as a binary vector $K^{(p)} \in \mathbb{R}^C$ using one-hot coding, where C is the number of ICU types. The sequence of heterogeneous events contains D numerical variables of length T that reflects the patient's physiological state. We denote it as $X^{(p)} \in \mathbb{R}^{T\times D}$, where $X_{t,d}^{(p)}$ $(t = 1,2,\dots,T, d = 1,2,\dots,D)$ represents the observations of d^{th} variable in the t^{th} time step. *Survival_days* denotes the number of days the patient survived between ICU admission and death.

Furthermore, the medical data inevitably carries missing observations. We introduce a masking matrix $M \in \mathbb{R}^{T\times D}$ to denote which variables are missing in the sequence of heterogeneous events. Specifically, we define

$$M(t,d) = \begin{cases} 1, & if\ X_{t,d}^{(p)}\ is\ observed \\ 0, & otherwise \end{cases}. \quad (1)$$

3.2 Problem Definition

Definition 1 (Survived): We define a patient is survived if the number of *survival days* between ICU admission and death is over 30.

Mortality Prediction Task: The mortality prediction is a time series classification problem. We predict whether the patient p is survived, given the *ICUType* $K^{(p)}$ and the sequence of heterogeneous events $X^{(p)}$.

4 P-BiLSTM Method

As shown in Figure 1, P-BiLSTM is composed of two modules, including a *Knowledge Representation* (KR) module that will be discussed in more details in Section 4.1 and a *Prediction Module* that will be discussed in Section 4.2. The proposed system works as follows. First, *KR* module extracts features from prior medical knowledge (e.g., *ICUType*). Subsequently, medical knowledge, heterogeneous clinical events and its labeling matrix are integrated as the input of ***weighted block***. After that, a shortcut between the output of weighted

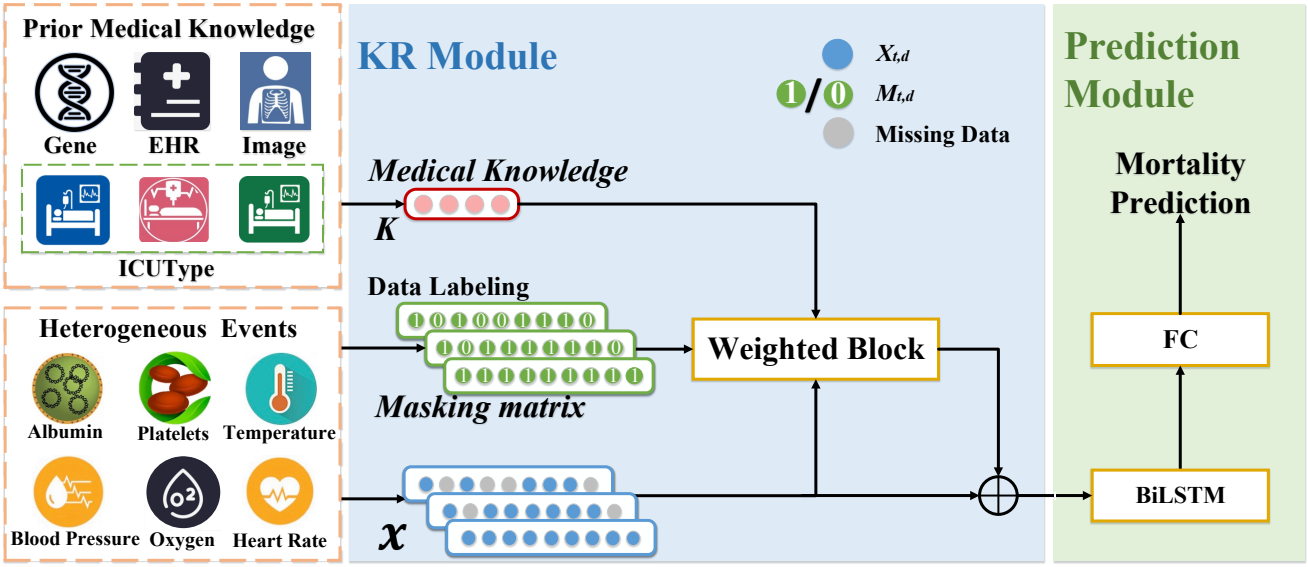

Figure 1: Overview of P-BiLSTM model. *KR module* extracts features from prior medical knowledge (e.g., *ICUType*), and *Prediction module* incorporated the aforementioned features into BiLSTM networks to predict ICU mortality.

block and heterogeneous clinical events are generated as features. Finally, *Prediction module* incorporated the aforementioned features into BiLSTM networks to predict ICU mortality.

4.1 Knowledge Representation (KR) Module

Besides sequences of heterogeneous clinical events, each patient's record contains pattern of missing sequential data and a group of important general descriptions, such as *ICUType*. In real world cases, data missing patterns and general descriptions are largely associated with the complex situation of diagnostic procedures, such as measuring blood pressure hourly, while collecting urine every 8 hours. To avoid noises introduced by straightforward data completion, it is necessary to incorporate indirect supervision (i.e., prior medical knowledge) with the deep neural networks to make prediction.

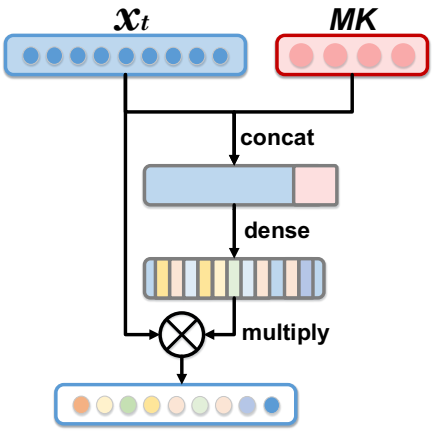

Figure 2: Weighted block. x_t is heterogeneous clinical events of time step t, MK is medical knowledge (e.g., *ICUType*)

Weighted Block

We design a ***weighted block*** to realize the importance of prior medical knowledge, which is illustrated in Figure 2. First, we collect relevant medical knowledge from experienced clinicians, such as causality that represents their various emphasis on different ICU types. Heterogeneous clinical events of time step *t*, $x_t \in \mathbb{R}^D$ are concatenated with the corresponding medical knowledge (i.e., $mk \in \mathbb{R}^N$) as a new vector $v_t \in \mathbb{R}^{D+N}$, which is then applied to a fully connected layer to generate the weight for each variable $w_t \in \mathbb{R}^D$. Finally, the input vector x_t is weighted by w_t, as

$$\hat{x}_t = x_t \otimes w_t, \quad (2)$$

where $\otimes$ denotes element-wise multiplication. The weighted block could guide the predictive model to predict decisions made by clinicians.

It is known that clinicians make diagnoses largely based on their considerations for a patient with a certain ICU type $K^{(p)} \in \mathbb{R}^C$, as well as a group of variables $X^{(p)} \in \mathbb{R}^{T \times D}$ accordingly. To reflect this in our design, as inspired by Residual network (ResNet) [He *et al.*, 2016], we add a shortcut connection between the sequence of heterogeneous events and outputs of the weighted block to improve prediction performance as illustrated in Figure 1.

Data Labeling Block

Medical datasets often carry missing observations. We observed that one causal factor of missing data relies largely on medical procedures. In other words, a straightforward data completion strategy that failed to consider medical routine procedure will inevitably introduce noises, which might interfere with the prediction model. For this reason, a better strategy is to label which data is completed explicitly to avoid noises in deep neural networks. In the paper, in order to label whether the variable is missing at each time step, we design a masking matrix $M^{(p)} \in \mathbb{R}^{T \times D}$. In Figure 1, the masking vector is used as an input to the weighted block in combination with sequences of heterogeneous events to reduce the impact of noises introduced by the missing data completion. For example, if a feature is absent, the normalized feature after processing will be penalized by the weighted block.

4.2 Prediction Module

Heterogeneous events consist of logic complex information. To obtain high mortality prediction accuracy, we take advantage of time-series data that capture heterogeneous events of patients to explore the pattern of the patient's physical condition changes.

Long short-term memory (LSTM) performs well in long-term time-series prediction problem. In our scenario, each patient's physical conditions are not only affected by the previous illness state, but also determined by present conditions. Bidirectional LSTM consists of forward and backward LSTMs, which helps us to avoid the blindness of unidirectional propagation and capture changes in patient's physical signs. Therefore, in the prediction module, we use bidirectional LSTM, and apply a fully connected prediction layer, which has sigmoid activation for the binary classification task (as shown in Prediction module in Figure 1).

5 Experiments and Evaluations

In this section, we first introduce the datasets and experimental settings, and then provide detailed performance comparison among the proposed P-BiLSTM and state-of-the-art approaches.

5.1 Datasets

PhysioNet Challenge 2012 dataset (*PhysioNet*)[1], is a publicly available collection of general descriptors and multivariate clinical time series from 4,000 ICU records. General descriptors in this dataset contain patients' basic information, including *recordID*, *age*, *gender*, *height*, *weight* and *ICUType*. *ICUType* specifies the type of ICU to which the patient has been admitted, including Coronary Care Unit, Cardiac Surgery Recovery Unit, Medical ICU, and Surgical ICU. The dataset also includes heterogeneous temporal events, which are composed of roughly 48 hours and contains

[1] PhysioNet website, https://www.physionet.org/challenge/2012/

37 variables that reflect each patient's physiological state such as *Albumin*, *heart-rate*, *glucose*, etc. More details of this dataset can be found on PhysioNet website.

5.2 Data Preprocessing

We used *PhysioNet* dataset in our experiment. We preprocessed the dataset in three categories. First, for general descriptors, we represented *ICUType* as a $K \in \mathbb{R}^{1\times4}$ boolean matrix by one-hot code as part of input data. Second, for heterogeneous events, the preprocessing steps included: *1) data recording*. We chose one hour as a time interval and statistic variable value of each sample in each time period with a $X \in \mathbb{R}^{48\times37}$ matrix. At the same time, we recorded whether the value is null with a $M \in \mathbb{R}^{48\times37}$ masking matrix; *2) data completion*. If a record is missing occasionally within 48 hours, we imputed data that are missing using neighboring records. If there is no data in 48 hours, missing data were replaced with the mean value of the variable in the same type of ICU the patient belongs to; and *3) data normalization.* To make heterogeneous temporal events comparable, the matrix uses the mean and standard deviation for normalization. Third, for patients' labels, we labeled each patient as {0, 1}, according to the survival definition described in Section 3.2.

5.3 Experimental Settings

In order to balance positive and negative samples, we used up-sampling method to expand 4,000 records of *PhysioNet Challenge 2012 dataset* to 5,000 samples in random order. The dataset was split into three parts with fixed proportions, namely training set (3360 samples), validation set (840 samples), and testing set (800 samples). Besides, algorithms were implemented using TensorFlow and Keras, and experiments were run on two GPUs (GTX TITAN X).

Model	*Precision*	*Recall*	*F1*	*AUC*
CNN	0.675	0.650	0.682	0.804
GRU	0.745	0.828	0.784	0.883
LSTM	0.755	0.853	0.801	0.905
BiGRU	0.757	0.848	0.800	0.894
BiLSTM	0.779	0.845	0.826	0.906
BiLSTM-ST	0.751	0.749	0.750	0.864
P-BiLSTM	**0.842**	**0.857**	**0.849**	**0.923**

Table 1: Performance of mortality prediction tasks

Comparing Methods

The following models were compared with ***P-BiLSTM***: (1) Convolutional Neural Network (**CNN**) [Krizhevsky *et al.*, 2015]; (2) Gated Recurrent Unit (**GRU**) [Chung *et al.*, 2014] (3) Long short-term memory (**LSTM**) [Hochreiter and Schmidhuber, 1997]; (4) Bidirectional GRU (**BiGRU**) [Yang *et al.*, 2016]; (5) Bidirectional LSTM (**BiLSTM**) [Graves *et al.*, 2005]; and (6) Bidirectional LSTM network with supervision technique (**BiLSTM-ST**) [Zhu *et al.*, 2018].

Evaluating Metrics

We choose four widely used metrics, i.e., *Precision*, *Recall*, *F1*, and the area under ROC Curve (*AUC*) to compare the performances of our model against baseline methods.

5.4 Results Summary

Comparison with Baselines

Table 1 shows that the proposed P-BiLSTM outperforms the-state-of-the-art methods. Specifically, P-BiLSTM outperforms BiLSTM, mainly because the prior medical knowledge could guide models to learn optimal parameters. In addition, the result of P-BiLSTM is statistically significant according to Student's T-test at level 0.063 compared to BiLSTM.

Ablation Studies

We re-trained our model by ablating two blocks to examine the effectiveness of our proposed techniques, namely the Weighted Block and Data Labeling Block. As shown in Table 2, the experimental results show that: 1) each block is useful for the prediction task; and 2) the full model that integrates with two blocks preforms the best.

Model	*Precision*	*Recall*	*F1*	*AUC*
W/O WB	0.815	0.856	0.835	0.914
W/O DLB	0.825	0.857	0.842	0.920
Full Model	**0.842**	**0.857**	**0.849**	**0.923**

Table 2: Ablation studies. *W/O WB* denotes no weighted block is performed, and *W/O DLB* denotes no data labeling block is performed.

Effect of Various Sequence Lengths

We further trained P-BiLSTM model with 24-hour and 36-hour heterogeneous temporal events after the patient was admitted into the ICU. Figure 3 shows that the performance of P-BiLSTM(36-hour) is slightly weaker than P-BiLSTM(48-hour). Nevertheless, we could still use 36-hour model instead of 48-hour so as to give clinicians an earlier sense of which patients will require critical targeted treatments.

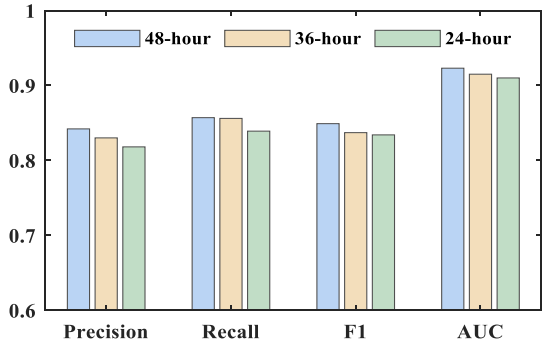

Figure 3: Performance with various sequence lengths.

6. Conclusion

In this paper, we propose a novel mortality prediction method P-BiLSTM for ICU, which integrates prior medical knowledge into deep neural networks to enhance the learning and explainable abilities. Specifically, we train and evaluate our model on a real-world dataset. Experimental results demonstrate that P-BiLSTM outperforms all other baseline methods. In the future, we plan to employ causality discovery technologies (i.e., do-Calculus) to enhance the interpretation of the mortality prediction method. In addition, we will testify our model in a large dataset (i.e., MIMIC III).

Acknowledgement

The work is supported by grants from the Natural Science Foundation of China (61872306). The corresponding authors is Xiaoliang Fan.

References

[Bhattacharya *et al.*, 2017] Sakyajit Bhattacharya, Vaibhav Rajan, and Harsh Shrivastava. ICU mortality prediction: A classification algorithm for imbalanced datasets. In *Thirty-First AAAI Conference on Artificial Intelligence*. pages 1288--1294, AAAI press, 2017. San Francisco.

[Chung *et al.*, 2018] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555,* 2014.

[Citi and Barbieri, 2012] Luca Citi and Riccardo Barbieri. PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. *Computing in Cardiology*, 25(1): 257–260, 2012.

[Graves *et al.*, 2015] Alex Graves, Santiago Fernández and Jürgen Schmidhuber Graves. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In *15th International Conference on Artificial Neural Networks*, pages 799--804, Springer Press, Warsaw, 2005.

[Gupta *et al.*, 2018] Priyanka Gupta, Pankaj Malhotra, Lovekesh Vig and Gautam Shroff. Using Features from Pre-trained TimeNet for Clinical Predictions. *The 3rd International Workshop on Knowledge Discovery in Healthcare Data at IJCAI*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770-778, 2016.

[He *et al*., 2019] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30-36, 2019.

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

[Johnson *et al.*, 2016] 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. MIMIC-III, a freely accessible critical care database. *Scientific data, 3(2016): 160035*, 2016.

[Krizhevsky *et al.*,2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *26th International Conference on Neural Information Processing Systems*, pages 1097--1105, MIT Press, Lake Tahoe, 2012.

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

[Liu *et al.*, 2018] Luchen Liu, Jianhao Shen, Ming Zhang, Zichang Wang, and Jian Tang. Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Ma *et al.*, 2018] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018.

[Nguyen *et al.*, 2017] Phuoc Nguyen, Truyen Tran, and Svetha Venkatesh. Deep learning to attend to risk in ICU. In *2nd International Workshop on Knowledge Discovery in Healthcare Data*, pages: 25–29, Morgan Kaufmann Press, 2017. Melbourne.

[Page *et al*., 2018] David Page, Enyo Ablordeppey, Brian T. Wessman, Nicholas M. Mohr, Stephen Trzeciak, Marin H. Kollef, Brian W. Roberts and Brian M. Fuller. Emergency department hyperoxia is associated with increased mortality in mechanically ventilated patients: a cohort study. *Critical Care*, 22(1): 9, 2018.

[Suo *et al.*, 2017] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. Personalized Disease Prediction Using A CNN-Based Similarity Learning Method. In *Proceedings of The IEEE International Conference on Bioinformatics and Biomedicine*, pages 811–816, 2017.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480-1489, 2016.

[Zhu *et al*., 2018] Yao Zhu, Xiaoliang Fan, Jinzhun Wu, Xiao Liu, Jia Shi, Cheng Wang, Predicting ICU Mortality by Supervised Bidirectional LSTM Networks, *In Proceedings of 1st Joint Workshop on AI in Health collocated with IJCAI 2018 (IJCAI-AIH 2018)*, pages 49-60, 2018.