DeepDiagnosis: DNN-based Diagnosis Prediction from Pediatric Big Healthcare Data

Jia Shi[1], Xiaoliang Fan[2,5,1*], Jinzhun Wu[3], Jian Chen[4], Wenbo Chen[1*]
[1] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
[2] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China
[3] The First Affiliated Hospital, Xiamen University, Xiamen, China
[4] ZOE Soft Co. Ltd., Xiamen, China
[5] Digital Fujian Institute of Healthcare & Biomedical Big Data Research, Xiamen University, Xiamen, China
Email: shij2016@lzu.edu.cn, fanxiaoliang@xmu.edu.cn, 1923731201@qq.com, chenjian@zoesoft.com.cn, chenwb@lzu.edu.cn

***Abstract*—Mining electronic health records (EHRs) has been considered as a major decision-making tool for clinical diagnosis. In fact, it is difficult to extract the valuable information from EHRs due to free-text writing, incomplete description, and high variabilities of diseases. Especially for pediatric EHRs, the shortage of experienced pediatricians as well as complex environmental factors such as seasonal variations, cross infections from kindergartens, make it extremely challenging to conduct a precise diagnosis. To address those challenges, we proposed *DeepDiagnosis*, a novel deep neural network-based diagnosis prediction algorithm by mining massive pediatric EHRs. First, we pre-process the unstructured EHRs dataset in Chinese and transfer them into sentence vectors by natural language processing technologies. Second, we construct the bidirectional recurrent neural networks (BiRNN) model to catch the patients' clinical symptoms as well as their interaction. Finally, we train and evaluate our model using a real-world dataset containing 81,476 pediatric EHRs. Experimental results show that the proposed method outperforms many baseline methods.**

Keywords— electronic health records, diagnosis prediction, deep neural networks.

I. INTRODUCTION

The global health care systems are rapidly adopting electronic health records (EHRs [1]), which are systematic collections of longitudinal patient health information (e.g., diagnosis, medication, lab tests, procedures, etc.). EHRs [1, 2], describing the symptoms and treatment of patients, can be a useful guidance and decision supports for those inexperienced clinicians. Furthermore, mining EHRs can also help researchers to study the regulation and characteristics of infectious diseases.

In fact, it is not easy to extract the valuable information for diagnosis prediction from massive EHRs, the main reasons include [3]: 1) free-text writing of clinicians; 2) incomplete description in EHRs; and 3) high variabilities of diseases. Especially for pediatric EHRs [4]. It is even difficult to precisely conduct a diagnosis prediction. For one thing, there is a huge shortage of experienced pediatricians that pediatric EHRs are of low quality in general. For another thing, environmental factors such as seasonal variations, cross infections from kindergartens, are playing an important role in the evolution of pediatric diseases.

Inspired by studies on electronic health records with deep neural networks (DNNs) [1, 6, 7, 15], we aim to make the best use of the strong learning ability of DNNs to represent the clinical symptoms hidden in the electronic health records. However, it is a non-trivial task to use EHRs for diagnosis prediction, and we need to address the following three challenges. First, it is challenging to distinguish respiratory system diseases because of subtle differences in symptoms. Fig. 1 shows that the number of outpatients in pediatric department is likely to be a long-tail distribution, where respiratory system diseases form the majority of all diseases. For example, the top four International Classification of Diseases - 10th version (ICD-10) code account for 49.84% of total outpatient number, and they are acute upper respiratory infections (J06.900), acute bronchitis (J20.900), bronchitis (J40.x00), bronchopneumonia (J18.000), respectively. However, there are very subtle differences among respiratory system diseases in terms of symptoms. Thus, it brings an obstacle for classification.

Second, when it comes to writing EHRs in a clinical scenario, there are variances among clinicians' free-text writings on the description of the same symptom. For instance, "coughed for 7 days" and "started to cough one week ago" actually describe the same symptom but in different writings. In addition, we are ought to make extra efforts to deal with EHRs standardization by natural language processing technologies, as our EHRs are written in Chinese.

Third, the pediatric diseases could be easily influenced by complex environmental factors such as seasonal variations, cross infections from kindergartens, etc. For instance, at the beginning of a new term, there will usually be a pandemic flu outbreak. As a result, those factors might lead to the sudden outbreak of pediatric diseases, such as many respiratory system diseases.

*Correspondence author.
[1] Electronic health record, https://en.wikipedia.org/wiki/Electronic_health_record

978-1-5386-8034-6/18/$31.00 ©2018 IEEE
DOI 10.1109/CBD.2018.00058

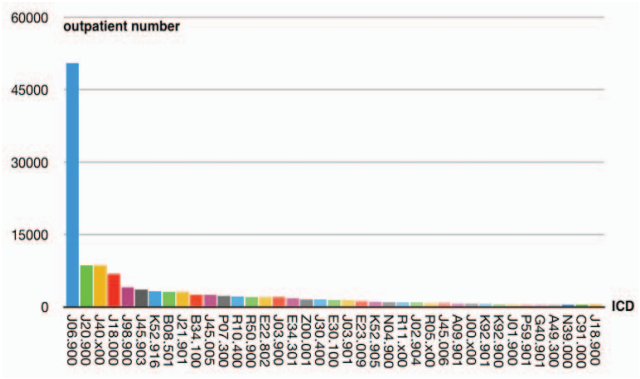

Figure 1. the distribution of outpatients number by ICD-10 in pediatric department from Jan. 2016 to Sept. 2017

To address those challenges, we proposed *DeepDiagnosis*, a novel deep neural network-based diagnosis prediction algorithm by mining massive pediatric EHRs. First, we preprocess the unstructured EHRs data in Chinese and transfer them into sentence vectors by natural language processing technology. Second, we construct the BiRNN model to uncover the complex interactions among patients' clinical symptoms. Finally, we train and evaluate the proposed method using a real-world dataset containing 81,476 EHRs from a pediatric department in a major city of China. Experimental results show that the proposed method outperforms many baseline methods.

The remainder of this paper is organized as follows. Section II introduces the related works. Section III presents basic notations and describes the observations of the datasets. Section IV proposes the detailed method. Section V demonstrates the experimental settings and the evaluation results. Finally, Section VI summaries this paper and points out the future work.

II. Related Works

Electronic Health Records (EHRs) are systematic collections of longitudinal patient health information [1]. Mining EHRs data is considered as a vital step for diagnosis prediction in hospitals and its main applications including electronic genotyping and phenotyping [2, 3, 5], disease progression [5-7], adverse drug event detection [8], diagnosis prediction [9-11], etc. For pediatric EHRs [4], it is difficult to precisely make a diagnosis prediction due to the shortage of experienced pediatricians as well as complex environmental factors such as seasonal variations, cross infections from kindergartens.

Many early works have investigated the diagnosis prediction from EHRs with machine learning and statistical techniques such as logistic regression, support vector machines (SVM), and random forests [5]. More recently, deep learning [13] models are gradually utilized to capture the complex and long-range dependencies in EHRs. Recurrent neural networks (RNNs) can be used for modeling multivariate time series data in healthcare with missing values [4,11]. Convolutional neural networks (CNNs) are used to predict unplanned readmission and risk [12] with EHRs. Stacked denoising autoencoders (SDAs) are employed to detect the characteristic patterns of physiology in clinical time series data [14]. Other DNN-based diagnosis prediction methods are Unsupervised Pre-trained Networks, Recursive Neural Networks, Long Short-Term Memory Networks.

Although our outpatient EHRs are not real time series data, we find that in pediatric department, most patients suffer from respiratory system diseases, which usually happened in outbreak and cross infection. Meaning that, both the past and the future context of the sequence have a relationship with the current records to a certain extent. Which is on the same principle as RNN.

III. Preliminary and Datasets

A. Basic Notations

In this work, we assume that there are N patients, the n-th patient has $T(n)$ visit records in the EHR data. The patients can be represented by a sequence of visits $V_1, V_2, \dots, V_{T(n)}$. Each visit Vt, contains a medical code describing the disease in International Classification of Diseases (ICD-10) and clinical narratives representing the patient's clinical symptoms in free text. We set the clinical narratives as the model's input, and medical code as label, respectively.

B. Datasets

We retrieved the daily EHRs data from a hospital in Xiamen, which varies from 0~14 years old patients from January 2016 to September 2017. Each patient's diagnosis is marked by International Classification of Diseases, tenth edition (ICD-10). Totally, there are 149,817 records of outpatient visits (41% females), generating 81,476 EHRs. The records for each patient consist of two parts: general descriptors and clinical information. First, general descriptors contain patients' basic information, including patient's ID, patient's name, age, gender. Second, clinical information is composed of 37 characters which records their clinical symptoms and treatments. For instance, parameters in Table I are Visit_ID, Visit_Time, Patient_ID, Patient_Name, Patient Gender, Patient_Age, Diagnosis record in ICD-10 code by the clinicians, etc.

Throughout the paper, we report only average statistics of the dataset without revealing any identifiable information of individuals. In addition, the EHRs dataset used in this work was authorized by the hospital's institutional review board.

TABLE I. EHRs Sample Records

Name	Sample
Visit_ID	6987173XXX
Visit_Time	2016/1/3 16:49:00
Patient_ID	3003149XXX
Patient_Name	First_Name, Family_Name
Patient_Gender	Male
Patient_Age	3
Diagnosis record (ICD-10)	J06.900

C. *Observations*

1) *General Descriptors*

We perform the basic statistics on general descriptors including patients' basic information. Fig. 2 shows the preliminary observations: 1) male patients are slightly more than female patients; and 2) children under 3 years old take the highest proportion (52%).

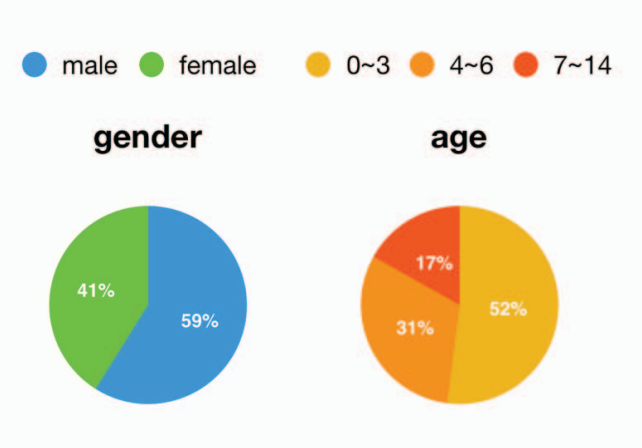

Figure 2. General statistics of patients.

Based on these observations, we could have the following findings: 1) the distribution of gender is generally balanced in samples; and 2) children in 0 to 3 years old are easier to fall ill than the rest of children.

2) *Influence Factors*

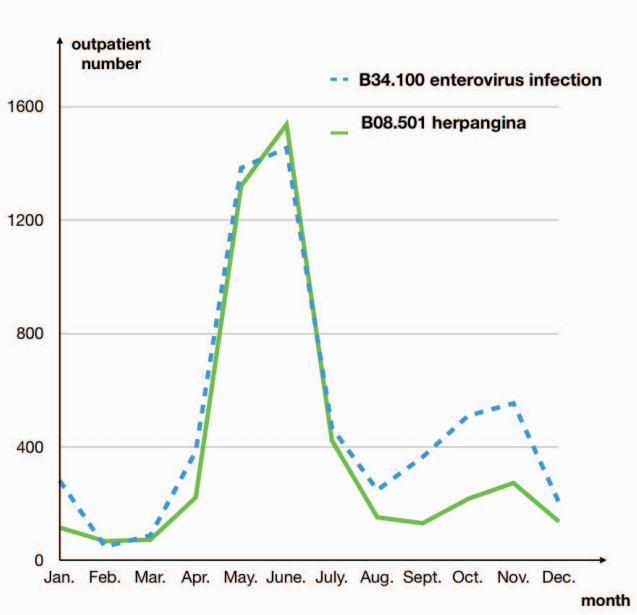

Figure 3. outpatients' number of herpangina and enterovirus infection in 2016

It is widely known that the pediatric diseases could be easily influenced by complex environmental factors such as seasonal variations, cross infections from kindergartens, etc. First, children are more sensitive than adults in term of seasonal variations. As shown in Fig. 3, *herpangina* and *enterovirus infection* have a large number of outpatients in May and June. For example, *herpangina*, also known as mouth blisters, is the name of a painful mouth infection caused by the *coxsackie* virus, which is a virus usually bred in the late spring and early summer. Suggesting that, a certain type of diseases might outbreak seasonally on children.

Second, children are in poor resistance, making them more easily affected by cross infections than adults. We observe in Fig. 4 that factors of social events, such as new term kick-off, could make a great impact on the number of outpatients. For example, *acute bronchitis*, also known as a chest cold, which is a short-term inflammation of the bronchi (large and medium-sized airways) of the lungs, usually outbreaks in spring due to the dry weather and the bloom of the viruses. However, as shown in Fig. 4, on Feb. 7th, 2016, there is a trough as most people went to hometown for Chinese New Year holiday. On the contrary, there is a rapid rise on March 6, 2016, due to cross infections caused by the beginning of the new semester in kindergartens or primary schools. In general, social events could play an important role in disease outbreaks that we should take it into account in our diagnosis prediction model.

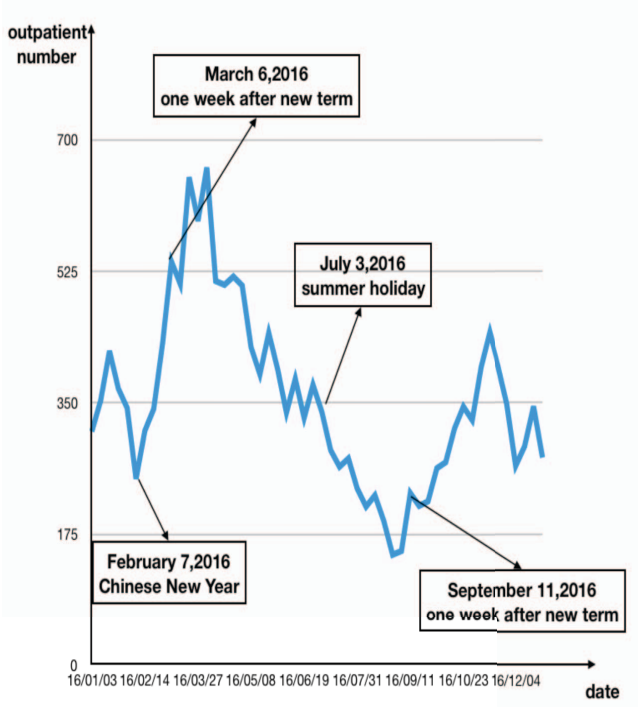

Figure 4. outpatients' number of acute bronchitis in 2016

IV. The Proposed Method

A. Framework

The overall framework of diagnosis prediction model is presented in Fig. 5. First, we preprocess the dataset to get the high quality EHRs data and extract useful clinically information from EHRs by natural language processing. Second, the processed data are used as input data for the bidirectional model, and we will get a prediction by the classifier. Finally, the loss function is calculated by comparing the predicted value with the true label to train model parameters.

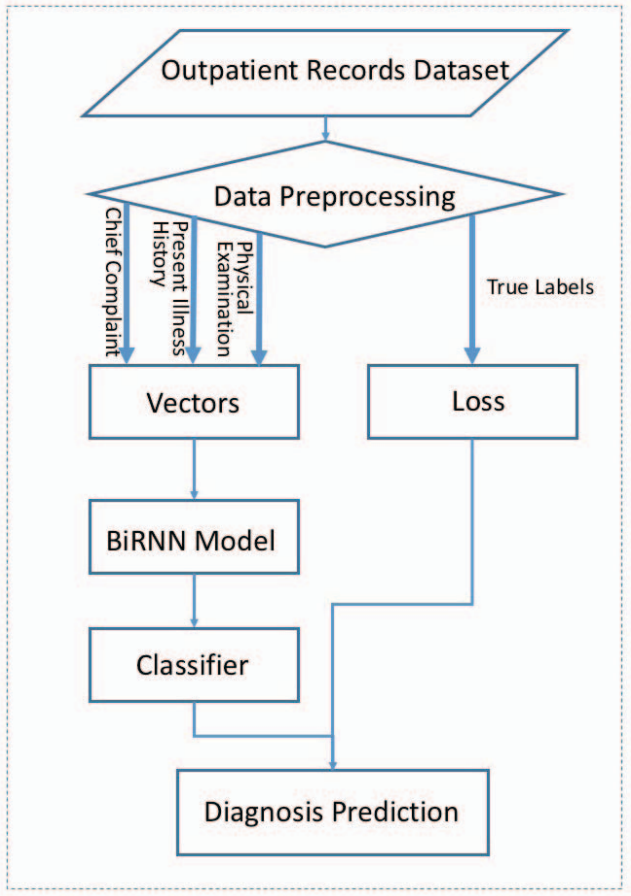

Figure 5. Overview framework

B. Data Preprocessing

There are three steps for data preprocessing. First, we choose 8 diseases which are most common in the pediatric department. In addition, we select 3 key features: chief complain, present illness, and physical examination.

Second, we removed the incomplete and incorrect data. In the outpatient department, when a patient require a laboratory test only, there is an outpatient record without a diagnosis code which is interference to our experiment. We also exclude those records which is lack of key information such as clinical symptoms. In addition, we removed inevasible bias such as systematic errors in the free-text records, which might result in significant bias when EHRs data are used naively for clinical research.

Third, we employed natural language processing to transfer EHRs data to vectors which will later be trained for our prediction model. EHRs are usually written in an unstructured manner, therefore it is necessary to translate the free-text information into vectors. Specifically, we use the Word2Vec tool to retrieve the vector form of each words. Then, we build the vectors of the free-text sentences. Finally, we join all the vectors of sentences together to represent a patient's clinical features.

C. Algorithm

In order to realize this classification model to make the diagnosis prediction, we employ a deep learning method to construct the model. Specially, we use BiRNN to extract the relationship among the outpatient electronical health records.

BiRNN is employed to learn temporal dependencies between the past and future frames. A fully connected layer is used to gather the outputs and learn a sequence representation followed by an 8-class softmax layer for classification.

Originally, RNN is a deep neural network with memory and deep in time which is developed for modeling dependencies in time series. Therefore, RNN model is suitable for our classification task with the advantage of encoding contextual information for sequences [16, 17]. We employ a BiRNN to simultaneously capture forward and backward dynamic transforms of sequences, i.e., two RNNs are respectively used to traverse the temporal sequence in a forward or backward behavior. The structure of BiRNN with two hidden layers and its expansion of full neutral network are shown in Fig. 6.

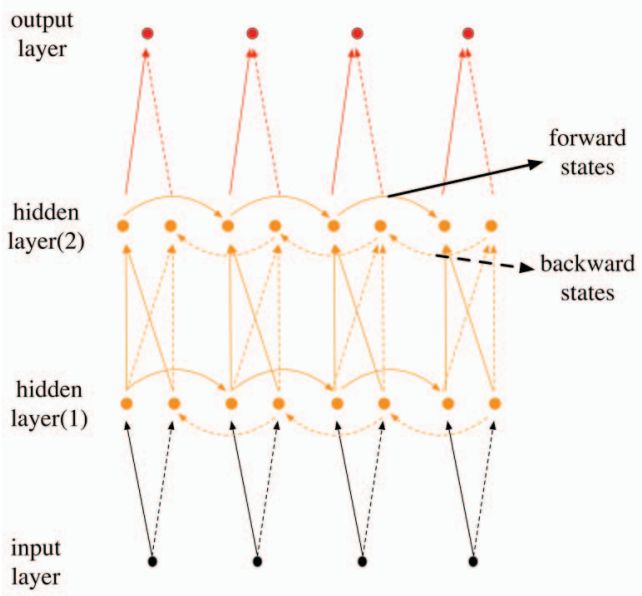

Figure 6. Structure of bidirectional recurrent neural network (BiRNN).

Suppose that sequential representations are denoted as x and the length is T, then the temporal BiRNN layer can be written as follows.

$$h_t^f = f_h\left(W_{hh}^f h_{t-1}^f + W_{ih}^f x_t^f + b^f\right) \quad (1)$$

$$h_t^b = f_h\left(W_{hh}^b h_{t-1}^b + W_{ih}^b x_t^b + b^b\right) \quad (2)$$

$$o_t = V^f h_t^f + V^b h_t^b \quad (3)$$

$$x_t^b = x_{T-t+1}^f \quad (4)$$

where $\{W_{hh}^f, W_{ih}^b, b^f\}$ and $\{W_{hh}^b, W_{ih}^b, b^b\}$ are the learned weight matrices parameters for recurrently traversing the sequences scanned forward and backward respectively. x_t^b, h_t^f, h_t^b, o_t are respectively denoted as the input nodes, hidden nodes for the forward scanned network, hidden nodes for the backward scanned network, and the corresponding output nodes at the state $t \in [1, \dots, T]$. b^b and b^f are the bias terms and f_h are defined as the non-linear activation function. Finally, the output nodes of BiRNN layer, denoted as $O = [o_1, \dots, o_t, \dots, o_T]$, are fed into softmax layer for emotion classification. We utilize the fully connected layers to collect the all the hidden unit outputs in the BiRNN layers, and then connect them with the final sequence labels. One fully connected layer and a softmax layer is applied in our network.

$$y = f_o(W_{ho} O + b_o) \quad (5)$$

$$O = [o_1, \dots, o_t, \dots, o_T] \quad (6)$$

where W_{ho} is the fully connected layer weight matrices. O is the concatenation of all its sequential states $o_t, t \in [1, \dots, T]$, b_o is the bias terms, f_o is softmax, the predicted class label is y.

As the last output node memorizes this necessary emotion information while adaptively forgetting those external noises in the evolutional of sequence network. Thus, we only choose the last output of state T, o_T as the final sequence characteristic, i.e.,

$$P(i|X) = exp(o_{Ti}) \,/\, \sum_{k=1}^{C} exp(o_{Tk}) \quad (7)$$

where $P(i|X)$ represents the probability for the input X being predicted as the i^{th} classes. In the objective function, the cross entropy loss is defined, which can be represented as:

$$L = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_i^c \, log \, P(c|I^i) \quad (8)$$

where L denotes the cross entropy loss, N is the number of training samples, I^i denotes the i^{th} training samples of the training set which is denoted as I , and y_i denotes the label of the i^{th} training sample.

V. Experiments and Evaluations

In this section, we will describe the detailed implementation of the experiment, including the experimental settings, evaluation metrics, baselines and results analysis.

A. Experimental Settings

We perform the experiment on 81,476 cases of the dataset, with 65,180 (80%) patients as the training set and the remaining 16,296 (20%) as the testing set. We apply the supervision technique to three different types of recurrent neural networks including CNN, RNN, LSTM, BiLSTM, BiRNN, so as to produce a variety of models for pediatric diagnosis prediction for comparison purpose. There are 32 cells per hidden layer in our final model, and the softmax function is utilized to classify the output into 8 categories.

The algorithms are developed using Python 3.6.1, TensorFlow 1.2.1 and Keras 2.1.1. Meanwhile, the experiment is conducted on 14 CPU cores (Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00 GHz), with two GPUs (GeForce GTX TITAN X).

B. Evaluation Metrics

As the prediction of mortality is a classification problem, we choose *Precision* and correct number of the result (C#) to evaluate our model and compare with baselines.

C. Baselines

We compare our BiRNN model with the following four baselines:

- *CNN*: Convolutional neural network is one of the most popular models of deep learning but it lacks handling of time series data.
- *RNN*: Recurrent neural network is a class of artificial neural network where connections between units form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.
- *LSTM*: Long short-term memory network is a special kind of RNN. It is explicitly designed to avoid the long-term dependency problem and it is capable of remembering information for long periods of time.
- *BiLSTM*: Bidirectional long short-term memory network has both forward and backward LSTM structure and it has a better overall understanding of time series data.

D. Results Summary

As shown in Table II, the BiRNN model outperforms baselines in general. BiRNN(2) stands for a BiRNN model with two hidden layers.

TABLE II. Comparison among different methods.

Model	*Precision*	*C#*
CNN	80.397	12564
RNN	79.608	12524
LSTM	80.409	12650
BiLSTM	80.556	12673
BiRNN(2)	**80.912**	**12729**
BiRNN(3)	80.289	12631

First, BiRNN performs better than RNN, and BiLSTM performs better than LSTM. As bidirectional model can use a finite sequence to predict or label each element of the sequence based on both the past and the future context of the element,

indicating that the bidirectional model could capture the complex and long-range dependencies related to the disease outbreak. Second, BiRNN is better than BiLSTM. The reason is that LSTM could perform well for long time series data rather than capturing the complex interaction among different patients in a short time.

VI. Conclusion

In this paper, we proposed a novel deep learning based diagnosis prediction framework. We trained and tested our model with a real-world dataset of 81,476 patients to catch the patients' clinical symptoms, using BiRNN to model the complex dependencies in sequential representation of patients. Our experimental results demonstrated that the BiRNN model could outperform those baseline methods.

In the future, our work can be extended, for example: 1) more sophisticated natural language preprocessing steps will be conducted to extract valuable information in EHRs data; and 2) more extensive outpatient EHRs datasets will be employed to evaluate and improve our model.

Acknowledgement

The work is supported by grants from the Natural Science Foundation of China (61300232); the China Postdoc Foundation (2015M580564); and Fundamental Research Funds for the Central Universities (lzujbky-2016-br04). The corresponding authors are Xiaoliang Fan and Wenbo Chen. This work was done when Jia Shi did an research internship at Xiamen University.

References

[1] Y. Cheng, F. Wang, P. Zhang, J.Hu. "Risk Prediction with Electronic Health Records: A Deep Learning Approach," in Proceedings of Siam International Conference on Data Mining, 2016, pp. 432-440, doi: 10.1137/1.9781611974348.49.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Translational genetics: Mining electronic health records: towards better research applications and clinical care," Nature Reviews - Genetics, vol. 13, Jun. 2012, pp. 395–405, doi: 10.1038/nrg3208.

[3] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research," IMIA Yearbook of Medical Informatics Methods Inf Med, vol. 47, 2008, pp. 128-144.

[4] C.Daymont, M.E.Ross, L.A.Russell, A.G.Fiks, R.C.Wasserman, R.W.Grundmeier, Daymont, C, et al. "Automated identification of implausible values in growth data from pediatric electronic health records," J Am Med Inform Assoc, vol.24.6, Nov. 2017, pp. 1080-1087, doi: 10.1093/jamia/ocx037.

[5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 1096–1103, doi: 10.1145/1390156.1390294.

[6] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," Molecular systems biology, vol. 12, Jul. 2016, ,pp. 878, doi: 10.15252/msb.20156651.

[7] A. M. Sarroff and M. Casey, "Musical Audio Synthesis Using Autoencoding Neural Nets," Proceedings of the International Computer Music Conference, vol. September, Sep. 2014, pp. 14–20.

[8] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks," in Proceedings of Machine Learning for Healthcare Conference, 2016, pp. 301-318.

[9] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi, G. Burns, D. C. Classen, and D. W. Bates, "Medication-related Clinical Decision Support in Computerized Provider Order Entry Systems: A Review," Journal of the American Medical Informatics Association, vol. 14, Jan. - Feb. 2007, pp. 29–40, doi: 10.1197/jamia.M2170.

[10] L. A. Knake, M. Ahuja, E. L. McDonald, K. K. Ryckman, N. Weathers, "Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data," Bmc Pediatrics, vol.16, Apr. 2016,pp.59, doi: 10.1186/s12887-016-0592-z.

[11] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction," arXiv preprint arXiv: 1602.03686, 2016.

[12] P. B. Jensen, L. J. Jensen, and S. Brunak, "Translational genetics: Mining electronic health records: towards better research applications and clinical care," Nature Reviews - Genetics, vol. 13, Jun. 2012, pp. 395–405, doi: 10.1038/nrg3208.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.

[14] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific reports, vol. 6, May. 2016, pp. 26094, doi:10.1038/srep26094.

[15] A. Jagannatha, H. Yu, "Bidirectional Recurrent Neural Networks for Medical Event Detection in Electronic Health Records," Proceedings of the conference Association for Computational Linguistics North American Chapter Meeting, 2016, pp.473-482.

[16] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease Inference from Health-Related Questions via Sparsely Connected Deep Learning," Knowledge and Data Engineering, IEEE Transactions, vol. 27, Aug. 2015, pp. 2107–2119, doi: 10.1109/TKDE.2015.2399298.

[17] B.Shickel, P.J.Tighe, A.Bihorac, et al. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," IEEE Journal of Biomedical & Health Informatics, 2017, pp.1-1, doi: 10.1109/JBHI.2017.2767063.