Learning to Detect Pathogenic Microorganism of Community-acquired Pneumonia

Wenwei Liang
East China Normal University
51174500033@stu.ecnu.edu.cn

Wei Zhang[*]
East China Normal University
zhangwei.thu2011@gmail.com

Bo Jin
East China Normal University
824976094@qq.com

Jiangjiang Xu
Shanghai Children's Hospital
13621919395@163.com

Linhua Shu
Shanghai Children's Hospital
shulinhua@126.com

Hongyuan Zha
Georgia Institute of Technology
zha@cc.gatech.edu

ABSTRACT

Community-acquired pneumonia (CAP) is a major death cause for children, requiring an early administration of appropriate antibiotics to cure it. To achieve this, accurate detection of pathogenic microorganism is crucial, especially for reducing the abuse of antibiotics. Conventional gold standard detection methods are mainly etiology based, incurring high cost and labor intensity. Although recently electronic health records (EHRs) become prevalent and widely used, their power for automatically determining pathogenic microorganism has not been investigated. In this paper, we formulate a new problem for automatically detecting pathogenic microorganism of CAP by considering patient biomedical features from EHRs, including time-varying body temperatures and common laboratory measurements. We further develop a Patient Attention based Recurrent Neural Network (PA-RNN) model to fuse different patient features for detection. We conduct experiments on a real dataset, demonstrating utilizing electronic health records yields promising performance and PA-RNN outperforms several alternatives.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → *Neural networks*;

KEYWORDS

Community-acquired pneumonia; Pathogenic microorganism detection; Deep learning

ACM Reference Format:
Wenwei Liang, Wei Zhang, Bo Jin, Jiangjiang Xu, Linhua Shu, and Hongyuan Zha. 2018. Learning to Detect Pathogenic Microorganism of Community-acquired Pneumonia. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210112

1 INTRODUCTION

Community-acquired pneumonia (CAP) [13] refers to the lungs of patients infected when they are not in hospital. It has long been a major cause of morbidity and death, especially for children. As reported by the studies [12, 15], pneumonia is one of the top ranked diseases responsible for the deaths of children both in USA and China. Curing CAP largely requires an early administration of appropriate antibiotics [9]. Unfortunately, the issue of the abuse of antibiotics is very prevalent, especially in developing countries such as China [7], which seriously endangers human health.

Alleviating the above issue needs an accurate detection of pathogenic microorganism [13]. Pathogenic microorganism is a family of microorganisms which will cause human diseases. If the pathogenic microorganism of CAP can be precisely identified, clinicians are able to prescribe optimal antibiotics. Conventional gold-standard detection methods are mainly etiology based, including culture-based assays, polymerase-chain-reaction (PCR), etc. However, many of them need specialized equipment and reagents, and are labor and time intensive [4, 17], which limit their application only in major hospitals. Thus, there is an urgent need to develop intelligent and cost-effective methodologies to detect pathogenic microorganism of CAP using data which is easier to be acquired.

Recent progress in wide collection of electronic health records (EHRs) [8] applies the methodologies from artificial intelligence community to CAP. However, existing studies in this regard are somewhat limited and mainly aim at 1) predicting whether suspected patients have pneumonia [16] or 2) further judging the risk of patients with pneumonia [3]. Most of them have ignored to investigate the power of patient easy-to-acquire data from EHRs for automatically detecting pathogenic microorganism of CAP. In fact, it plays a great role in treating CAP children. In this paper, we formulate a new problem of utilizing pneumonia patients' multiple medical features from EHRs to identify their pathogenic microorganisms. To our best knowledge, none of previous studies has investigated this problem. The studied features include time-varying body temperature and some carefully selected clinical measurements which are easy to be acquired, such as white blood cell count from routine blood test (see Table 1 for details). Consequently, the central challenge is how to effectively fuse the above multiple types of features and construct an effective model for the problem.

To address the challenge, we develop a Patient Attention based Recurrent Neural Network (PA-RNN), which is capable of modeling sequential body temperatures and fusing multiple types of patient features. To be specific, PA-RNN first exploits the power

[*] Corresponding author

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210112

of recurrent neural network (RNN) to obtain a sequence of body temperature representations for different time steps. Meantime it constructs patient basic features which are carefully selected from EHRs. Afterwards, inspired by attention mechanism [1], PA-RNN provides a patient feature based attention to determine the importance of each time-varying temperature representation and further gains an integrated representation for a whole body temperature sequence. Finally, the model fuses the integrated representation with the representation of patient basic features for pathogenic microorganism detection.

In a nutshell, the major novelty of PA-RNN is that most previous studies which utilize RNN to model EHRs [2, 5, 10, 11, 14] focus on predicting targets at the next time step based on current hidden states of RNN. However, we obtain an integrated representation of body temperatures sequence through a novel patient feature based attention computation to all hidden states of RNN. We conduct comprehensive experiments on a real world dataset from a major hospital in China, indicating the benefit of fusing multiple types of features from EHRs for the studied problem, and demonstrating the effectiveness of PA-RNN over several alternative methods.

2 COMPUTATIONAL MODEL

2.1 Problem definition

Assume the CAP record set is denoted as $\mathcal{R} = \{\mathcal{R}^{tr}, \mathcal{R}^{te}\}$, where $\mathcal{R}^{tr}$ is used for training and $\mathcal{R}^{te}$ for testing. Each record $r \in \mathcal{R}$ can be expressed as, $r = \{u, X_u, y_u\}$, where u denotes the pneumonia patient in the record, X_u represents the patient time-varying body temperatures and other features from EHRs, and y_u corresponds to the class of pathogenic microorganism causing pneumonia (e.g., mycoplasmal pneumonia (MP), bacterial pneumonia (BP), and respiratory syncytial virus pneumonia (RSVP)). Based on the above denotations, we formally define the problem as below,

Problem 1 (**Pathogenic Microorganism Detection**). *Given a training set $\mathcal{R}^{tr}$ of CAP, the target is to learn a model $f : X_u \rightarrow y_u$ for each record $r \in \mathcal{R}^{tr}$, and further utilize the model to detect pathogenic microorganism of target records in a test set $\mathcal{R}^{te}$.*

2.2 Patient features

We introduce the selected features from patients' EHRs that could be utilized as indicators for determining the pathogenic microorganism of CAP. All the selected features shown in Table 1 are categorized into three groups: 1) body temperatures, 2) clinical measurements, and 3) demographics. Among them, the features in the latter two groups are selected based on chi-square test and the advices from clinicians. We do not provide the results of the test due to page space limitation and it not being the major focus in this paper.

Body temperatures. Fever is a common comorbidity of CAP, leading to anomalous variation of body temperatures. We consider this type of feature, hoping to reveal sequential characteristics and benefit the detection of pathogenic microorganism. The time interval between consecutive temperature measurements in our dataset is about 2 hours. If not stated, we adopt patients' body temperature of the first two days in hospital, which ensures the time cost of our detection method is less than the traditional detection methods such as PCR.

Table 1: Summary of selected features.

Feature	Description
Temp	Time-varying body temperatures
Chest X-ray	"1" denoting lobar and "0" for the rest
WBC	White blood cell count
NE_per	Neutrophil percentage
LYM_per	Lymphocyte percentage
CRP	C-reactive protein
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ALB	Albumin
Season	"1" for summer and antumn and "0" for the rest
Age	Patient age when admission

It is intuitive that the variation of body temperature in normal range could be informative for the detection. Following the suggestion from physicians, we adopt the min-max strategy to rescale all values of time-varying body temperatures. The minimum temperature is set to 37.2°C while the maximum to 40.0°C. Suppose $V_u \in \mathbb{R}^{\mathrm{m}}$ denotes the numerical sequence of body temperature for patient u, $V_{u,t}$ represents the value of t-th time step in the sequence, and m is the total count of time steps. Then we can define the formula as follows:

$$V_{u,t} = \begin{cases} 0 & V_{u,t} < 37.2°\mathrm{C} \\ \frac{V_{u,t} - 37.2°C}{40.0°C - 37.2°C} & 37.2°\mathrm{C} \leq V_{u,t} \leq 40.0°\mathrm{C} \\ 1 & V_{u,t} > 40.0°\mathrm{C}\ . \end{cases} \tag{1}$$

Clinical measurements. The selected medical features are mainly infectious indicators to pathogenic microorganism. For example, we find that CRP has a closer association with BP through chi-square test. It is worth noting that we only consider the first measurements of these features when designing PA-RNN for the following reasons. First, these features are not repeatedly measured for some patients. For example, the average of AST measurement for each patient is 1.19 in our dataset. Second, the average intervals between two consecutive measurements are usually more than 5 days and much larger than those for body temperatures. If we want to consider their sequential characteristics, the time cost of collecting those data will be very high, which is against our purpose of detecting pathogenic microorganism faster than etiology based methods.

Demographics: We adopt "Age" and "Season" to denote user demographics. On the one hand, patients in different ages might be infected by different types of pathogenic microorganism with different possibilities. For example, we find that pneumonia patients in the age of 6 to 14 are more likely to be infected by MP. On the other hand, pneumonia has seasonal characteristics. For example, the proportion of patients with MP in summer and autumn is about 20% higher than those in the other two seasons.

For ease of later model specification, we regard the combination of laboratory measurements and demographics as **patient basic features** and denote their corresponding value vector as S_u for patient u in record r. Finally we can get $X_u = \{V_u, S_u\}$.

2.3 Model specification

We take the record r mentioned above as an example to introduce PA-RNN. The basic framework of the model is shown in Figure 1.

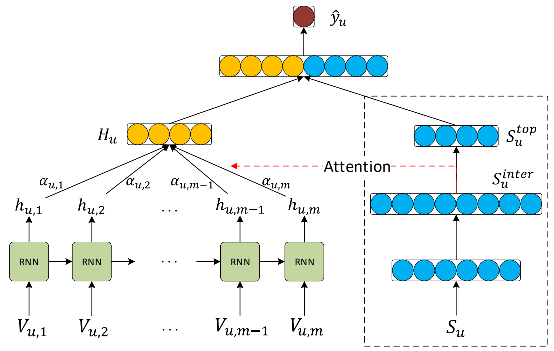

Figure 1: The architecture of the proposed PA-RNN model.

Overall, it consists of two essential components: 1) the left part of the figure presents a recurrent neural network for modeling time-varying body temperatures, and 2) the right part indicates a feed-forward neural network for modeling patient basic features. The two components are correlated in PA-RNN through the following two manners. First, PA-RNN utilizes the intermediate representation of patient basic features for attention calculation to obtain the integrated representation of body temperatures. Second, it concatenates the integrated representation and the final representation of patient basic features for pathogenic microorganism detection.

RNN for body temperatures: To model body temperatures, we adopt long short-term memory (LSTM) network [6] due to its good performance of modeling sequential data. It is capable of considering dependencies of temperatures between different time steps.

Given the input $V_{u,t}$ at time step t, the previous hidden state $h_{u,t-1}$, and the cell state $c_{u,t-1}$ of LSTM, we define the follow equations to obtain the current hidden state $h_{u,t}$,

$$\begin{bmatrix} i_{u,t} \\ f_{u,t} \\ o_{u,t} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \end{bmatrix} (W_{ifo} \cdot [V_{u,t}; h_{u,t-1}] + b_{ifo}), \tag{2}$$

$$c_{u,t} = i_{u,t} \odot \tanh(W_c \cdot [V_{u,t}; h_{u,t-1}] + b_c) + f_{u,t} \odot c_{u,t-1}, \tag{3}$$

$$h_{u,t} = o_{u,t} \odot \tanh(c_{u,t}), \tag{4}$$

where $i_{u,t}$, $f_{u,t}$, and $o_{u,t}$ correspond to the activations of input gate, forget gate and output gate, respectively. We use σ to denote the sigmoid function and $\odot$ to represent Hardmard product. W_{ifo}, W_c, b_{ifo}, and b_c are the learnable parameters of LSTM. After recurrent computation for each time step, we can obtain the hidden state sequence, $\{h_{u,1}, \ldots, h_{u,m}\}$.

Personalized attention computation: Before introducing the attention computation for the above hidden state sequence, we first define S_u^{inter} and S_u^{top} to denote the intermediate and top layers' outputs of the feed-forward neural network, respectively. Each layer of the network is associated with nonlinear activation functions, such as rectified linear unit (ReLU).

It is intuitive that temperatures in different time steps have different degrees of importance for representing the whole time-varying sequence, which will be used to detect pathogenic microorganism. We propose a novel attention computation to capture this intuition, which utilizes the intermediate representation of patient basic features to guide the computation of attention weights,

$$\alpha_{u,t} = \text{softmax}\Big(W_a \cdot \tanh(W_h h_{u,t} + W_s S_u^{inter} + b_a)\Big), \tag{5}$$

where W_a, W_h, W_s are weight matrices and b_a is a bias vector. Based on this, we can get the integrated representation H_u of body temperature sequence, i.e., $H_u = \sum_{t=1}^{m} \alpha_{u,t} h_{u,t}$.

Learning to detection. After getting H_u and S_u^{top}, we concatenate them to form a joint representation for patient u in the record r. With this representation, we could make the detection more accurate. Suppose the target is expressed as $\hat{y}_u$, then we can calculate it as follows:

$$\hat{y}_u = \text{softmax}(W_y \cdot [H_u; S_u^{top}] + b_y), \tag{6}$$

where W_y, b_y are the learnable parameters. We minimize PA-RNN by the cross entropy error between the real target y and the generated target $\hat{y}_u$ by gradient based methods.

3 EXPERIMENTS

3.1 Dataset and evaluation metrics

We study the problem of pathogenic microorganism detection for CAP using a real-world dataset from a Hospital in China, in which patients are all children but with different ages. Due to privacy issue, we anonymize all the patients. The EHRs were recorded from June 1st in 2014 to May 31st in 2015. The adopted patient features are already shown in Table 1. To handle missing values in patient basic features, we adopt the mean imputation strategy [10]. When the length of a patient time-varying body temperature is less than the pre-specified count of time steps m (e.g., 24), we use $37.2°C$ to pad temperature sequences from back to front, occupying about 30% of the dataset. It is reasonable because CAP patients are supposed to leave hospital when they are back to health with normal body temperature. In summary, we have 681 qualified records and each record corresponds to a unique patient. As about 48% of patients have MP, much larger than others like BP and RSVP (e.g., BP accounts for about 23%), we regard whether CAP patients having MP or not as the detection target.

Since the data size is not very large, we adopt 5-fold cross validation and report the average results. The evaluation metrics adopted in the experiments are average accuracy (Avg ACC) and average area under the curve (Avg AUC), which are commonly used in classification tasks.

3.2 Implementation details

We implemented our model and the other comparisons with the Keras library and Python. The Adam algorithm is adopted for training PA-RNN with a mini-batch size of 16 and the learning rate of 0.0005. L2 regularization is employed to alleviate the overfitting issue. All the methods are trained with maximum of 200 training epochs and the early stopping strategy is also considered. Without specific statement, the number of the time step m is set to 24 and the hidden state of LSTM is set to 5. The units of intermediate and top layers of feed-forward neural network are set to 24 and 16, respectively. To ensure fair comparison, we report the best performance for each method after tuning their hyper-parameters.

3.3 Comparison study

Comparison with alternatives. We choose the method of regarding the maximum class (MC) as the detection for its simplicity.

Table 2: Evaluation results of only modeling temperatures.

Method	Avg ACC	Avg AUC
MC	0.5213	0.5000
LR	0.5536	0.5473
GBDT	0.5624	0.5596
LSTM	**0.5800**	**0.6013**

Moreover, two standard classification models, linear logistic regression (LR) and nonlinear gradient boosting decision tree (GBDT), are adopted for comparison. To fuse the mentioned patient features, we first concatenate the body temperature sequence with other basic features and denote the corresponding methods with a suffix "(Seq)". As the large temperature feature dimension might influence the performance of classifiers, we adopt the average body temperature as the feature instead of using the whole sequence. The corresponding methods are suffixed by "(Avg)".

We first test the performance of all the adopted methods considering only body temperatures. Table 2 shows results of LSTM and other compared methods. MC performs worst among all the methods because it does not consider any patient feature. GBDT outperforms LR, showing that nonlinear modeling for temperature sequence is promising. LSTM performs best among all the methods, which shows the its advantage for modeling sequential data and supports our model choice.

Table 3 compares our approach with other alternatives on all features. Our final model PA-RNN improves all the other models, including the variant of our model, PA-RNN (w/o attention), which does not use the attention computation. We can also see PA-RNN (w/o attention) outperforms LR and GBDT significantly. All the above phenomenons show that the improvements of PA-RNN are not only from utilizing LSTM for modeling sequential temperatures, but are also caused by the proposed effective attention computation.

More analysis. Due to space limitation, we only report that the average accuracy of PA-RNN considering only patient basic features is 0.7371, which shows that the integration of body temperature and patient basic features is indeed beneficial.

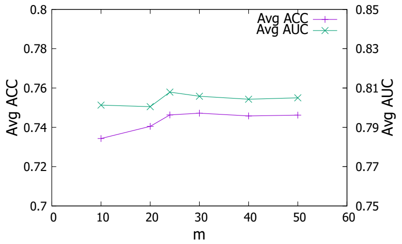

Figure 2: Results of different length of sequences.

We test how the results of PA-RNN differ with different length of temperature sequences. Figure 2 shows that when they become longer, the model achieves slightly better results. It is intuitive that longer sequences could bring more information about patients.

4 CONCLUSION

In this paper, we present a new problem of pathogenic microorganism detection for CAP patients by considering their features including time-varying body temperature from EHRs. We propose a deep learning model called PA-RNN with a novel attention computation, to model sequential body temperatures and fuse multiple types of features. Experimental results on a real world dataset prove the effectiveness of the proposed PA-RNN for pathogenic microorganism detection.

Table 3: Evaluation results of modeling all features.

Method	Avg ACC	Avg AUC
MC	0.5213	0.5000
GBDT (Seq)	0.7224	0.7211
LR (Seq)	0.7239	0.7232
GBDT (Avg)	0.7254	0.7236
LR (Avg)	0.7342	0.7326
PA-RNN (w/o attention)	0.7423	0.7974
PA-RNN (ours)	**0.7464**	**0.8079**

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. This work was supported in part by NSFC (61702190, U1609220), Shanghai Sailing Program (17YF1404500), Shanghai Chenguang Program (16CG24).

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015).

[2] Jacek M Bajor and Thomas A Lasko. 2017. Predicting Medications from Diagnostic Codes with Recurrent Neural Networks. (2017).

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *SIGKDD*. ACM, 1721–1730.

[4] Keping Chen, Runqing Jia, Li Li, Chuankun Yang, and Yan Shi. 2015. The aetiology of community associated pneumonia in children in Nanjing, China and aetiological patterns associated with age and season. *BMC public health* 15, 1 (2015), 113.

[5] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* 24, 2 (2016), 361–370.

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

[7] Mara Hvistendahl. 2012. China takes aim at rampant antibiotic resistance. *Science* 336, 6083 (2012), 795–795.

[8] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. 2009. Use of electronic health records in US hospitals. *New England Journal of Medicine* 360, 16 (2009), 1628–1638.

[9] Sushil K Kabra, Rakesh Lodha, and Ravindra M Pandey. 2010. Antibiotics for community acquired pneumonia in children. *Cochrane Database of Systematic Reviews* 3 (2010).

[10] Xiaohan Li, Shu Wu, and Liang Wang. 2017. Blood Pressure Prediction via Recurrent Models with Contextual Layer. In *WWW*. 685–693.

[11] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2015. Learning to diagnose with LSTM recurrent neural networks. *ICLR* (2015).

[12] Sherry L Murphy, Jiaquan Xu, and Kenneth D Kochanek. 2012. Deaths: preliminary data for 2010. *National vital statistics reports* 60, 4 (2012), 1–52.

[13] Daniel M Musher and Anna R Thorner. 2014. Community-acquired pneumonia. *New England Journal of Medicine* 371, 17 (2014), 1619–1628.

[14] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *PAKDD*. 30–41.

[15] Igor Rudan, Kit Yee Chan, Jian SF Zhang, Evropi Theodoratou, Xing Lin Feng, Joshua A Salomon, Joy E Lawn, Simon Cousens, Robert E Black, Yan Guo, et al. 2010. Causes of deaths in children younger than 5 years in China in 2008. *The Lancet* 375, 9720 (2010), 1083–1089.

[16] Insu Song. 2015. Diagnosis of pneumonia from sounds collected using low cost cell phones. In *IJCNN*. 1–8.

[17] Seung Min Yoo and Sang Yup Lee. 2016. Optical biosensors for the detection of pathogenic microorganisms. *Trends in biotechnology* 34, 1 (2016), 7–25.