ICIC Express Letters
Volume **14**, Number **12**, December **2020**
 ISSN 1881-803X
pp. **1155–1162**

REAL-TIME MEDICAL ELECTRONIC DATA MINING BASED HIERARCHICAL ATTENTION MECHANISM

Yi Mao[1], Yun Li[1] and Yixin Chen[2]

[1]School of Computer Science and Technology
Nanjing University of Posts and Telecommunications (NJUPT)
No. 9, Wenyuan Road, Yadong New District, Nanjing 210023, P. R. China
maoyi@njupt.edu.cn

[2]Department of Computer Science and Engineering
Washington University in St. Louis
One Brookings Drive, St. Louis, MO 63130, USA
chen@cse.wustl.edu

Received March 2020; accepted June 2020

Abstract. *Data mining on clinical data has great potential to improve the treatment quality of hospital and increase the survival rate of the patients. Data-driven prediction technology strongly hinges on the data collection and analysis of patients' vital signs. Deep neural networks are supported by a Recurrent Neural Network (RNN) architecture with Long Short-Term Memory (LSTM) units, and have achieved state-of-the-art results in a number of clinical prediction tasks. Recently, the architecture based on attention mechanism has achieved remarkable success in migration tasks, and has higher computing power in NLP (Natural Language Processing). In this paper, we recur to hierarchical attention and encoder-to-decoder based model to automatically learn features from medical records of time series of vital sign, categorical features which include demographics, hospitalization history, vital sign and laboratory tests. Moreover, instead of working as a black unexplainable box, we present the approach to extract potential informative risk factors, thereby helping doctors to make optimal decisions. Experiments show that our model is effective in extracting meaningful features, while the hierarchical attention mechanism can provide a better insight into relationships between different types of medical time series.*

Keywords: Deep learning, Medical electronic time series, Data mining, Attention mechanism

1. **Introduction.** Identifying disease and giving patient treatment timely, accurately and effectively will give them more opportunities for survival and protect their organ function maximally. Consolidating and analyzing large databases have the opportunity to transform the health-care industry. In this paper, we propose to develop and study an attention-based deep learning framework for existing hospital patient Electronic Health Records (EHR). Mining the aggregation of such multi-scale multi-source data can lead to novel tools to facilitate optimized patient-centered, evidence-based decision making can give the alert when the deterioration is happening or about to happen. Being able to detect these phases automatically would save many hours of time spent by entomologists to analyze the patient electronic data manually. A large amount of information in a time series is hidden in its structure, not only in numerical values. We propose a framework to effectively train deep architectures to learn hidden discriminant features from the original time series in an end-to-end manner.

Powered by Recurrent Neural Network (RNN) architectures with Long Short-Term Memory (LSTM) units, deep neural networks have achieved state-of-the-art results in

DOI: 10.24507/icicel.14.12.1155

several clinical prediction tasks [1, 2] since it is effective in exploiting long-range dependencies and handling nonlinear dynamics. Despite the success of RNNs, its sequential natural prohibits parallelized computing, thus making it inefficient particularly when processing long sequences [3]. Though many efforts have been done to improve the computational efficiency, some of the limitations still persist. Vaswani et al. [4] argued that attention mechanism can be effective in sequence-to-sequence modeling tasks without any recurrence. Also, attention mechanisms can be used to capture dependencies in sequences without considering their actual distances in the sequence [3].

Weakly-labeled training data may contain extraneous/irrelevant sections. The differences in the global time series are very subtle. It is very common for medical electronic data from Wireless Sensor Networks (WSN) to have distortions or invalid due to patient movement or sensor disconnections, and these are likely to confuse any global measures of time series. As local features, by introducing attention mechanism, we can be brittle to low level of noise and distortions. By leveraging different degrees of smoothness in compositional function, deep learning models show the ability to handle dimensional curses [5].

This paper is organized as follows. Section 2 describes the algorithms employed in the system, including RNN, attention mechanism. Section 3 introduces the related work in data mining in medical field and preliminary on attention mechanism. Section 4 presents the experimental results on several datasets. In Section 5, we form our conclusions.

2. **Proposed Model.** This section details the proposed clinical risk prediction model using deep learning for analyzing a large volume of multi dimensional heterogeneous clinical data. We design a hierarchical structure and two forms of attention for the decoder as shown in Figure 1. The architecture adopts the typical encoder-decoder structure. In the encoder, a CNN is used to extract features from each clinical time series (such as heart rate, temperature and blood pressure which are recorded every minute). LSTM then is used to capture the long and short dependency in the time series. In the decoding stage, decoding operates on a hierarchical structure with the hidden states of the LSTM. Finally, we employ a concatenation layer to combine the information from both vectors to get the diagnosis prediction.

For each particular patient, there is a corresponding piece of EHR, which consists of several time series. Assuming we use r variables, the records of the nth patient can be represented by a sequence of T^n tuples $\left(t_i^n, x_i^n \in \mathbb{R} \times \mathbb{R}^r,\ i = 1, \ldots, T^{(n)}\right)$. Here, t_i^n denotes the time of the ith visit of the nth patient. To minimize clutter, we describe the algorithms for a single patient and have dropped the superscript n whenever it is unambiguous. The goal of predictive modeling is to predict the label at each time step $y_i \in \{0, 1\}^s$, or at the end of the sequence $y \in \{0, 1\}^s$. The number of labels s can be more than one.

2.1. **Encoder.** In the encoder, a CNN with 1-dimensional kernel is used to extract features from time series. We use multiple layers of feature maps to convolute the data from raw-data level to feature level. In the case of multiple time series, the same convolution operation is applied to each one while keeping the output of the convolution separated since each time series has its own physical characteristics, and a convolution across different time series might negatively affect the extraction of the typical patterns for a specific one.

CNN features are generally not a good representation of series data. To get the series aspect of the clinical time series, we apply an LSTM to the features extracted by the CNN. For the sth time series, we will have hidden state of the LSTM, denoted by $R^s = \{r_1^s, r_2^s, \ldots, r_m^s\}$, where m is the number of the hidden states. To simplify the symbol, we use g_s to represent the final hidden state of the LSTM for the sth time series.

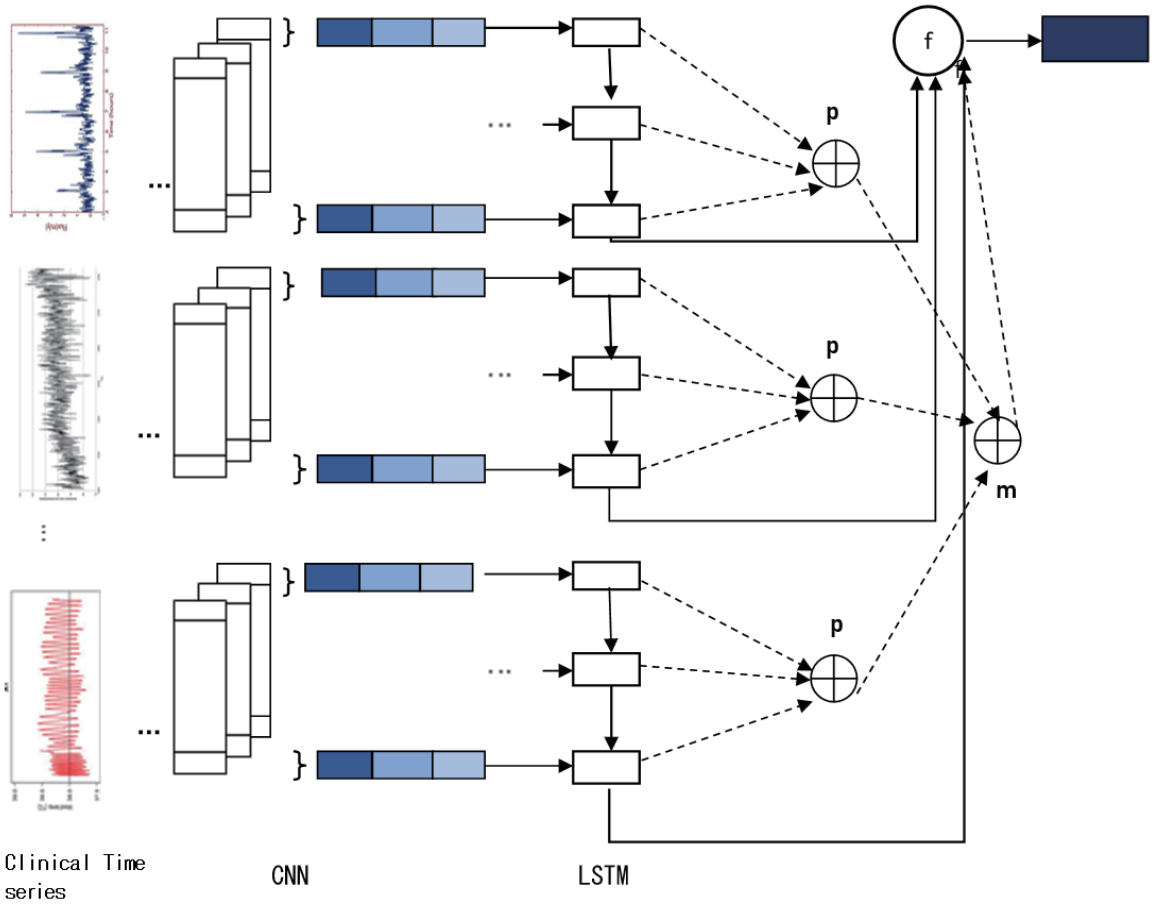

FIGURE 1. Architecture of the proposed method

By applying CNNs and LSTMs to all the time series, we can obtain a set of the final hidden states, which is denoted by $G = \{g_1, g_2, \ldots, g_{N_s}\}$, where N_s is the number of the time series.

2.2. **Hierarchical attention mechanism.**

2.2.1. *Attention inside time series.* Similarly to the work in [6], this form of attention is done inside time series, which retrieves information from relevant subsequences of a specific time series. The alignment is done as follows. Suppose that $R = \{r_1, r_2, \ldots, r_m\}$ is the set of the hidden states of LSTM for the series, and g_s is final hidden states. The similarity between g_s and r_i ($i \in \{1, 2, \ldots, m\}$) is calculated by

$$v_i = \mathbf{V}^T \tanh(W_1 \cdot g_s + W_2 \cdot r_i) \tag{1}$$

where $W_1, W_2 \in R^{(d \times d)}$, $\mathbf{V} \in R^d$. Then v_i is normalized as:

$$\alpha_i = \frac{\exp(v_i)}{\sum_{i'} \exp(v_i')} \tag{2}$$

By weightily averaging over all hidden states, vector $\mathbf{p}_t$ is calculated by

$$\mathbf{p}_t = \sum_{s \in [1, N_s]} \alpha_i r_i \tag{3}$$

2.2.2. *Attention between time series.* This form of attention is done between time series, which aligns the final hidden states of the LSTM in each time series and the label of the patient. The alignment is done as follows. Suppose that $P = p_1, p_2, \ldots, p_{N_S}$ is the set of

the final states in word level, and y_{t-1} is the label at the previous time step $t-1$. The similarity between y_{t-1} and p_s ($s \in 1, 2, \ldots, N_S$) is calculated by

$$v_s = \mathbf{V}^T \tanh(W_3 \cdot y_{t-1} + W_4 \cdot p_s) \tag{4}$$

where $W_3, W_4 \in R^{d \times d}$, $\mathbf{V} \in R^d$. Then v_s is normalized as:

$$\alpha_s = \frac{\exp(v_s)}{\sum_{s'} \exp(v_s{}')} \tag{5}$$

where α_s is the weight between $(t-1)$th state and the sth time series. It can be seen as a metric for measuring the importance of their relation. By weighted averaging over all types of time series, vector $\mathbf{m}_t$ is calculated by

$$\mathbf{m}_t = \sum_{s \in [1, N_s]} \alpha_s p_s \tag{6}$$

2.2.3. *Diagnosis prediction.* Given the context vector $\mathbf{m}_t$ and the hidden state R^s, we employ a simple concatenation layer to combine the information from both vectors to generate an attention hidden state as follows:

$$\tilde{h_t} = \tanh\left(\mathbf{W}_c \left[\mathbf{m}_t; R^s\right]\right) \tag{7}$$

where $\mathbf{W}_c$ is the weight matrix. The attentional vector $\tilde{h_t}$ is fed through the softmax layer to produce the $(t+1)$th visit information as defined as:

$$\tilde{y_t} = Softmax\left(\mathbf{W}_s \tilde{h_t} + \mathbf{b}_s\right) \tag{8}$$

2.2.4. *Objective function.* Here, we use the cross-entropy between the ground truth visit information y_t and the predicted label $\tilde{y_t}$ to calculate the loss for all the patients as follows:

$$L(x_1, \ldots, x_T) = -\frac{1}{N} \sum_{t=1}^{N} \left(y_t^\top \log\left(\tilde{y_t}\right) + (1 - y_t)^\top \log\left(1 - \tilde{y_t}\right)\right) \tag{9}$$

2.3. **Interpretation.** In healthcare, the interpretability of the learned representations of medical codes and visits is important. We need to understand the clinical meaning of each dimension of medical code representations and analyze which one is critical to the prediction. Since the proposed model is based on attention mechanisms, it is easy to find the importance of each visit for prediction by analyzing the attention scores. For the ith prediction, if the attention score α_i^j is large, then the probability of the $(j+1)$ information related to the current prediction is high. We employ the simple method proposed in [7] to interpret the code representations. First, we use $\mathrm{ReLU}\left(W_v^T\right)$, a non-negative matrix to represent the medical codes. Then we rank the codes by values in a reverse order for each dimension of the hidden state vector. Finally, the top k codes with the largest values are selected. By analyzing the selected medical codes, we can obtain the clinical interpretation of each dimension.

3. **Related Work.** Data mining is another important field of artificial intelligence in medical field. Medical data has the characteristics of diversity, complexity, redundancy, timeliness and non-normatively. Closely the doctor's experience and the traditional statistical analysis cannot get the hidden rules in the data. Methodological support is provided by extracting implicit information. Early warning system based on data mining cannot limit the amount of data, expert experience, high-speed computing of computers, information processing, issue mining, the establishment of intelligent early warning and auxiliary diagnosis system. A large bunch of work currently exists designed to tackle these challenges. Linear dynamical system models the linear transition between consecutive states in time, and can be augmented by Gaussian Process (GP) to provide more general non-linear modeling on local sequences to deal with the irregular sampling

issue [8]. Ung et al. [9] used instantaneous state of heart rate to predict atrial fibrillation. McManus et al. [10] applied time-varying coherence function for atrial fibrillation detection. Marozas et al. [11] used fusion of irregularity of RR intervals (the time elapsing between two consecutive R waves in the electrocardiogram) and bigeminy suppression for Atrial Fibrillation (AF) detection. Lake and Moorman [12] extracted entropy as feature in very short physiological time series for atrial fibrillation detection. Chen et al. employed multi-scale convolutional neural networks for time series in [13]. Mao et al. [14] proposed an integrated data mining approach with the multi-statistical features. Somanchi et al. [15] extracted features from heterogeneous data source and employed Support Vector Machine (SVM) as classifier for cardiac arrest early prediction. Kim et al. [16] used extra physiological variables extracted from an APACHE crital care system. Almayyan [17] selected discriminative features using Particle Swarm Optimization (PSO) and several selection techniques to reduce the features dimension. Futoma et al. applied ANN for predicting early hospital readmission [18]. Wang et al. [19] proposed a cost-sensitive based multilayer perceptron to predict early readmission prediction. In order to deal with the multi-variate nature of measurements, Ghassemi et al. [20] proposed using Gaussian process to transform the records into specific latent space.

Deep architecture with a greater number of layers shows that deep learning can extract abstract and invariant features for better performance of EHR classification [1]. Deepr learns how to extract features with CNN with word embedding and max pooling from medical records [21]. Stacked Denoising Autoencoders (SDA) are used to capture regularities and dependencies in EHR to generate robust patient descriptors used to predict future patient diseases in [22]. Using SDA deep patient [23] can extract the hierarchical features and pattern from EHRs data. However, the lack of analyzing the meaning of the feature makes the model unexplainable. However, heterogeneous property of EHR data remains one of the key challenges to be addressed [24] and the advantages of evolving deep learning techniques have not yet been fully utilized. Besides this, most of the representations are learned in an unsupervised manner, and cannot promise that the reconstructed feature representations by these deep learning models can finally be useful to supervised tasks [25].

Aiming at learning representations that preserve spatial, spectral and temporal patterns, Recurrent Neural Networks (RNN) have been used to model EEG data [26]. In 2015, Phung et al. proposed to use LSTMs with additional training strategies for diagnosis tasks, so as DeepCare [27]. In 2016, Che et al. also introduced RNNs to automatically deal with missing values [28]. By joint training on all tasks in MIMIC-III datasets, RNN modeling have been further improved [1]. Doctor AI [29] utilizes sequences of pairs occurring in each patient's timeline across multiple admissions as input to a GRU network to forecast future diagnosis and medical prescriptions. Deep Belief Network (DBN) is also introduced in classifying patients from normal ones in clinical [30]. By working with Convolutional Neural Network (CNN), Generative Adversarial Networks (GAN) can also provide plausible labeled EHR data by mimicking real patient records, to augment the training dataset in a semi-supervised learning manner [31]. Although deep learning models can produce accurate predictions, these models are mainly treated as black-box models that lack interpretability and transparency of their inner working [32], which makes the clinical warning unreliable. The attention-mechanism based learning is recent trend [33] for understanding what part of historical information weights more in predicting. The original attention mechanism proposed in [34] aims at improving the performance of neural machine translation. When introduced to EHR modeling, attention weights can indicate the degree to which clinical events the model can predict disease onsets or future events.

4. Experiment.

4.1. **Data description.** To evaluate the proposed method, three datasets are adopted, including the Long-Term AF Database (LTAFDB), the MIT-BIH AF Database (AFDB) and one real dataset from Barnes-Jewish Hospital. LTAFDB and AFDB are public datasets, which could be accessible from [35]. The real dataset is from Washington University School of Medicine and Barnes-Jewish Hospital, one of the largest hospitals in the United States. The database is from the General Hospital Wards (GHWs) between July 2007 and July 2011.

For general hospital wards, 41,305 patient visits are involved and 2,565 have the outcome of readmission or not. In this dataset, each patient is measured for 34 indicators, including demographics, vital signs (pulse, shock index, mean arterial blood pressure, temperature, and respiratory rate), and lab tests (albumin, bilirubin, BUN, creatinine, sodium, apotassium, glucose, hemoglobin, white cell count, INR, and other routine chemistry and hematology results).

4.2. **Results and discussion.** For comparison, we follow the way of dataset division in [9]. The LTAFDB database is used as the training set to determine the model parameter, while AFDB is used as the testing sets. Table 1 lists the results on AFDB with six existing methods, such as Ung et al. [9], McManus et al. [10], Marozas et al. [11], Lake and Moorman [12], Chen et al. [13]. For RNN, we feed the embedding to GRU and use the hidden state produced by GRU to get the predicted result. For simplicity, we use Zhou15, Lee13, Petrenase15, Lake11, MCNN16, RNN, for short to represent the benchmark approaches. From the result, we could see our proposed hierarchical attention mechanism based method could improve the detection performance, the accuracy is 98.19%, while the sensitivity is 98.22% and the specificity is 98.68%.

Table 1. Prediction result on AFDB

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
Lake11	N/A	0.9400	0.9100	0.4200	0.7400	0.8600	0.7700
Lee13	0.9791	0.9768	0.9822	0.6203	0.7840	0.8634	0.7025
Petrenase15	N/A	0.9710	0.9830	0.5235	0.7400	0.8900	0.8100
Zhou15	0.9799	0.9844	0.9783	0.6615	0.7726	0.8614	0.7530
MCNN16	0.9818	0.9811	0.9822	0.6620	0.7832	0.8725	0.7243
RNN	0.9721	0.9536	0.9642	0.5244	0.7622	0.8316	0.7323
Our method	0.9819	0.9868	0.9822	0.6735	0.8044	0.8825	0.8125

For real dataset, we evaluate our method for comparison with existing approaches used in hospitals, such as Mao et al. [14], Somanchi et al. [15], Kim et al. [16], Almayyan [17], Futoma et al. [18], Wang et al. [19]. For RNN, we also feed the embedding to GRU and use the hidden state produced by GRU to get the predicted result. For simplicity, we use MaoKDD12, SomanchiKDD15, KimHIR14, Almayyan16, Futoma15, Haishuai17, RNN, for short to represent the benchmark approaches. Table 2 presents the performance of the different predictive approaches. In comparison to the state-of-the-art baselines on the test set, we find our model performs significantly better than those traditional feature based method, such as SomanchiKDD15, MaoKDD12, KimHIR14. Compared to other neural networks such as convolutional neural network and recurrent neural network, our method has higher PPV (36%) and AUC (0.70), which means the system creates less false alarms with high accuracy in predicting.

Besides these, the attention mechanisms make the whole system an explainable system which can extract potential informative risk factors and risk time series, thereby helping doctors to make optimal decisions. The attention map could tell which clinical time series

TABLE 2. Readmission prediction on the GHWs dataset

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
SomanchiKDD15	0.83	0.85	0.08	0.15	0.53	0.88	0.19
MaoKDD12	0.72	0.86	0.18	0.30	0.52	0.85	0.20
KimHIR14	0.85	0.85	N/A	0.00	0.61	0.89	0.08
Almayyan16	0.84	0.85	0.11	0.19	0.57	0.86	0.15
Futoma15	0.84	0.86	0.23	0.36	0.62	0.87	0.16
Haishuai17	0.87	0.89	0.27	0.41	0.69	0.83	0.35
RNN	0.87	0.88	0.25	0.32	0.66	0.83	0.33
Our method	0.87	0.9	0.28	0.40	0.70	0.88	0.36

plays important roles in causing the alarms and also could be an reference for doctors in further intervention.

5. **Conclusion.** In this paper, we have introduced a new framework for real-time medical stream data classification. Data mining on clinical data not only increases the quality of health care but also decreases medical expanses. In this paper, we aim to deliver superior prediction quality, with good interpretability and high computational efficiency. Taken as a whole, our results demonstrate the promise and widespread applicability of the hierarchical attention mechanism-based classification for stream data. This way can utilize the advantage of local information, temporal and global trends of vital medical stream data. Moreover, attention mechanisms can extract potential informative risk factors and risk time series period, which can help doctors to make optimal and further decisions.

Acknowledgment. This work is supported by NUPTSF (Grant No. NY217136). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] H. Harutyunyan, H. Khachatrian, D. C. Kale et al., Multitask learning and benchmarking with clinical time series data, *Scientific Data*, 2019.

[2] Y. Omae, M. Mori, T. Akiduki and H. Takahashi, A novel deep learning optimization algorithm for human motions anomaly detection, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.199-208, 2019.

[3] H. Song, D. Rajan, J. J. Thiagarajan and A. Spanias, Attend and diagnose: Clinical time series analysis using attention models, *AAAI*, pp.201-213, 2017.

[4] A. Vaswani, N. Shazeer, N. Parmar et al., Attention is all you need, *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp.6000-6010, 2017.

[5] E. Keogh and A. Mueen, Curse of dimensionality, *Encyclopedia of Machine Learning*, pp.257-258, 2011.

[6] J. Li, M. T. Luong and D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp.1106-1115, 2015.

[7] E. Choi, M. T. Bahadori, E. Searles et al., Multi-layer representation learning for medical concepts, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pp.1495-1504, 2016.

[8] Z. Liu and M. Hauskrecht, Learning adaptive forecasting models from irregularly sampled multivariate clinical data, *Proc. of the AAAI Conference on Artificial Intelligence*, pp.1273-1279, 2016.

[9] B. Ung, E. Pickwell-MacPherson, X. Zhou, H. Ding and Y. Zhang, Automatic online detection of atrial fibrillation detection algorithm based on the instantaneous state of heart rate, *PloS One*, vol.10, 2015.

[10] D. D. McManus, J. Lee, Y. Nam and K. H. Chon, Time-varying coherence function for atrial fibrillation detection, *IEEE Trans. Biomedical Engineering*, vol.60, no.10, pp.2783-2793, 2013.

[11] V. Marozas, A. Petrenas and L. Sornmo, Low-complexity detection of atrial fibrillation in continuous long-term monitoring, *Computers in Biology and Medicine*, vol.65, pp.184-191, 2015.

[12] D. E. Lake and J. R. Moorman, Accurate estimation of entropy in very short physiological time series: The problem of atrial fibrillation detection in implanted ventricular devices, *AJP Heart and Circulatory Physiology*, vol.300, pp.319-325, 2011.
[13] Y. Chen, Z. Cui and W. Chen, Multi-scale convolutional neural networks for time series classification, *arXiv:1603.06995*, 2016.
[14] Y. Mao, W. Chen, Y. Chen et al., An integrated data mining approach to real-time clinical monitoring and deterioration warning, *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp.1140-1148, 2012.
[15] S. Somanchi, S. Adhikari, A. Lin, E. Eneva and R. Ghani, Early prediction of cardiac arrest (code blue) using electronic medical records, *Proc. of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pp.2119-2126, 2015.
[16] S. Kim, W. Kim and R. W. Park, A comparison of intensive care unit motality prediction models through the use of data mining techiques, *Healthcare Informatics Research*, vol.17, no.4, pp.232-243, 2011.
[17] W. Almayyan, Lymph disease prediction using random forest and particle swarm optimization, *Journal of Intelligent Learning System and Applications*, 2016.
[18] J. Futoma, J. Morris and J. Lucas. A comparison of models for predicting early hospital readmissions, *Journal of Biomedical Informatics*, 2015.
[19] H. Wang, Z. Cui, Y. Chen et al., Cost-sensitive deep learning for early readmission prediction at a major hospital, *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, 2017.
[20] M. Ghassemi, M. A. Pimentel, T. Naumann et al., A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data, *AAAI*, pp.446-453, 2015.
[21] P. Nguyen, T. Tran, N. Wickramasinghe and S. Venkatesh, Deepr: A convolutional net for medical records, *IEEE J. Biomedical and Health Informatics*, vol.21, no.1, pp.22-30, 2017.
[22] R. Miotto, L. Li and J. T. Dudley, Deep learning to predict patient future diseases from the electronic health records, *European Conference on Information Retrieval*, pp.768-774, 2016.
[23] 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*, 2016.
[24] H. Li et al., Identifying informative risk factors and predicting bone disease progression via deep belief networks, *Methods*, vol.69, no.3, pp.257-265, 2014.
[25] Z. Huang, W. Dong, H. Duan and J. Liu, A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records, *IEEE Trans. Biomedical Engineering*, vol.65, no.5, 2018.
[26] P. Bashivan, I. Rish, M. Yeasin and N. Codella, Learning representations from EEG with deep recurrent-convolutional neural networks, *arXiv:1511.06448*, 2015.
[27] D. Phung, T. Pham, T. Tran and S. Venkatesh, DeepCare: A deep dynamic memory model for predictive medicine, in *Advances in Knowledge Discovery and Data Mining. PAKDD 2016. Lecture Notes in Computer Science*, J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Huang and R. Wang (eds.), Cham, Springer International Publishing, 2016.
[28] Z. Che, S. Purushotham, K. Cho et al., Recurrent neural networks for multivariate time series with missing values, *Scientific Reports*, pp.6085-6097, 2018.
[29] E. Choi, M. T. Bahadori, A. Schuetz et al., Doctor AI: Predicting clinical events via recurrent neural networks, *Machine Learning and Healthcare Conference (MLHC 2016)*, Los Angeles, CA, 2016.
[30] M. M. Rahhal et al., Deep learning approach for active classification of electrocardiogram signals, *Inf. Sci.*, vol.345, pp.340-354, 2016.
[31] Z. Che, Y. Cheng, S. Zhai et al., Boosting deep learning risk prediction with generative adversarial networks for electronic health records, *ICDM 2017*, 2017.
[32] Lipton and Z. Chase, The mythos of model interpretability, *Communications of the ACM*, vol.61, no.10, 2016.
[33] F. Ma, R. Chitta, J. Zhou et al., Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, *arXiv:1706.05764*, 2017.
[34] Y. Bengio, D. Bahdanau and K. Cho, Neural machine translation by jointly learning to align and translate, *ICLR 2015*, 2015.
[35] X. Zhou, H. Ding, W. Wu and Y. Zhang, A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate, *PloS One*, vol.10, 2015.