

# Early Prediction of Sepsis From Clinical Data via Heterogeneous Event Aggregation

Luchen Liu\*, Haoxian Wu\*, Zichang Wang, Zequan Liu, Ming Zhang

Department of Computer Science, Peking University, Beijing, China

## Abstract

*Sepsis is a life-threatening condition that seriously endangers millions of people over the world. Hopefully, with the widespread availability of electronic health records (EHR), predictive models that can effectively deal with clinical sequential data increase the possibility to predict sepsis and take early preventive treatment. However, the early prediction is challenging because patients' sequential data in EHR contains temporal interactions of multiple clinical events. And capturing temporal interactions in the long event sequence is hard for traditional LSTM. Rather than directly applying the LSTM model to the event sequences, our proposed model firstly aggregates heterogeneous clinical events in a short period and then captures temporal interactions of the aggregated representations with LSTM. Our proposed Heterogeneous Event Aggregation can not only shorten the length of clinical event sequence but also help to retain temporal interactions of both categorical and numerical features of clinical events in the multiple heads of the aggregation representations. In the PhysioNet/Computing in Cardiology Challenge 2019 [1], with the team named PKU\_DLIB, our proposed model, in high efficiency, achieved utility score (0.321) in the full test set.*

## 1. Introduction

Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its tissues and organs. And the early prediction of the sepsis onset is important for physicians to take early preventive treatment. However, sepsis prediction is a difficult task, because there are complex sepsis risk factors including the age of patient, immune system weakness, complication (e.g. cancer, diabetes), conditions (e.g. trauma, burns) and so on. Hopefully, with the help of the widespread availability of electronic health records (EHR), well-designed predictive models, which can effectively make use of clinical sequential data, will be able to increase the sepsis prediction per-

formance.

The early sepsis prediction is challenging because patients' sequential data in EHR contains temporal interactions of multiple clinical events [2, 3]. The interactions of multiple clinical events include event co-occurrence in a short period (e.g. two related symptoms occur together) and event temporal dependency at large time-scale (e.g. A vital signal abnormally arises several hours after certain drug injection). One possible solution is directly applying deep sequential models, such as LSTM [4], Transformer[5], on the clinical event sequence. However, capturing temporal interactions in the long event sequence is hard for traditional LSTM because the length of clinical sequences exceeds the modeling ability of LSTM.

Rather than directly applying the LSTM model to the event sequences, some works design hierarchical neural networks to model the long sequence[6]. For example, aggregating events in a short period into a vector helps to shorten the original long sequence[7]. However, the information of each kind of clinical events is mixed in the aggregation vector, so temporal interactions of these events are hard to capture.

To address these issues, our proposed model firstly aggregates heterogeneous clinical events in a short period and then captures temporal interactions of the aggregated representations with LSTM. The Heterogeneous Event Aggregation module can not only shorten the length of clinical event sequence but also help to retain temporal interactions of both categorical and numerical features of clinical events in the multiple heads of the aggregation representations. The separated clinical information in different heads makes it easier to capture event temporal interactions in different aggregation vectors. Experiments on the PhysioNet/Computing in Cardiology Challenge 2019 show that our proposed model is effective and efficient compared to traditional methods. The contributions of this work is summarised as following:

- We propose a model to capture temporal interactions among multiple types of clinical event streams from EHR data for early sepsis prediction.
- The proposed heterogeneous event aggregation module can reduce the length of long clinical event sequences and

\* The two authors have equal contribution to this work.

retain their temporal interactions.

- Our proposed model achieves good prediction performance and time efficiency.

## 2. Dataset and Preprocessing

### 2.1. Dataset

The EHR data provided publicly for this challenge is sourced from two separate ICU, containing 20000 and 20643 records respectively. Each record is made up of hourly clinical data for a specific patient. Each row represents a single hour’s data with 40 variables and an additional label indicating whether the patient will get sepsis within 6 hours. With a positive sample proportion of 7.21%, there are 2932 sepsis patients in total. Sepsis and normal patients are divided into train and test set at the same ratio respectively. 5-fold cross validation is established over the train set.

### 2.2. Data Preprocessing

In this competition, our goal is to make an early sepsis detection within 6 hours for every time-point without causal model. For a patient’s record, We use a fix-length sliding window, with 1 hour step, to sample fix-length records  $x_t = r_{t-L+1:t}$  (zero filling if  $t < L$ ).

The 40 columns of the record contain 37 numerical variables and 3 binary variables. At the preprocessing stage, we relabel the 3 variables of two-categorical from 0 to 6 (6 clinical categories and one empty category for NaN). Finally we reorder the categorical variables to the last columns. As for numerical variables, in order to make the deep model converge easily, for each numerical variable, z-score normalization is applied.

## 3. Proposed Model

In this section, Heterogeneous Event Aggregation (HEA) module is designed to effectively capture the interaction information among the heterogeneous clinical events. The motivations of HEA are listed following: (1) Modeling interaction of both categorical and numerical heterogeneous clinical events from their embedding. (2) Grouping events into multiple heads in different aspects. (3) Shortening the length of clinical event sequence.

After the feature is extracted from HEA, the temporary dependency is captured by bidirectional LSTM. At last, the final outputs of 2 direction is sum up and forwarded to a single dense layer with sigmoid activation to get the detection.

The proposed architecture is shown in Figure 1. Given a sequential clinical record  $(X_1, X_2, X_3 \dots X_L)$   $X_i \in \mathbb{R}^{40}$ , our objective is to generate an early prediction of sepsis

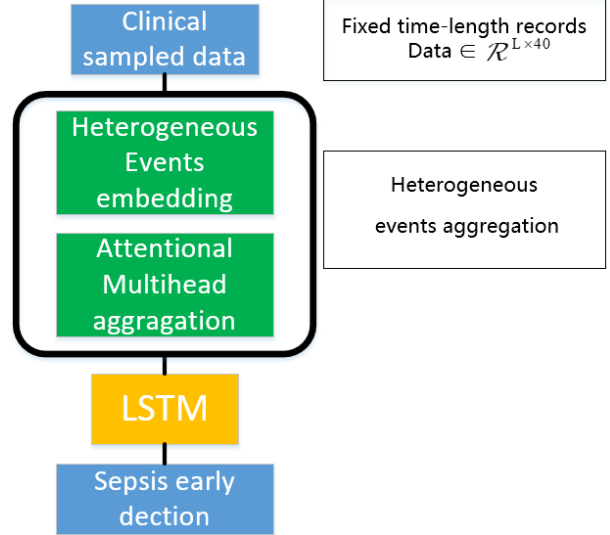


Figure 1. The architecture of our proposed model

for the last time step  $X_L$ . Heterogeneous Event Aggregation (HEA) module is composed of two parts, Heterogeneous Events Embedding and Attentional Multihead Aggregation, which are specifically explained as follows.

### 3.1. Heterogeneous Events Embedding

Given the sequential clinical data, the first step of our model is to generate the embedding that can be used to capture the interaction representation among the heterogeneous clinical events [8]. For each time step  $X_t \in \mathbb{R}^{40}$  (the first 37 columns are numerical variables, and the last 3 are categorical variables). Randomly initialized numerical event vector book  $W_{ne} \in \mathbb{R}^{37 \times d}$ , categorical event lookup table  $W_{ce} \in \mathbb{R}^{7 \times d}$  and value vector table  $W_{vn} \in \mathbb{R}^{37 \times d}$  are generated. The embedding for  $X_t$  is then generated as:

$$E_t = Mask(Concat(En_t, Ec_t)) \quad (1)$$

$$En_t = W_{ne} + X_t[:37]W_{vn} \quad (2)$$

$$Ec_t = Embedding\_lookup(X_t[37:], W_{ce}) \quad (3)$$

$$K_t = Concat(W_{ne}, Ec_t) \quad (4)$$

Mask function is used to mask the event embedding to zero if the corresponding variable is default.  $d$  is the dimensional numbers of embedding.  $E_t \in \mathbb{R}^{40 \times d}$  is combination of both numerical and categorical embedding,  $En_t \in \mathbb{R}^{37 \times d}$  is the numerical embedding,  $Ec_t \in \mathbb{R}^{3 \times d}$  is the categorical embedding,  $K_t \in \mathbb{R}^{40 \times d}$  is the Key matrix of both numerical and categorical events.

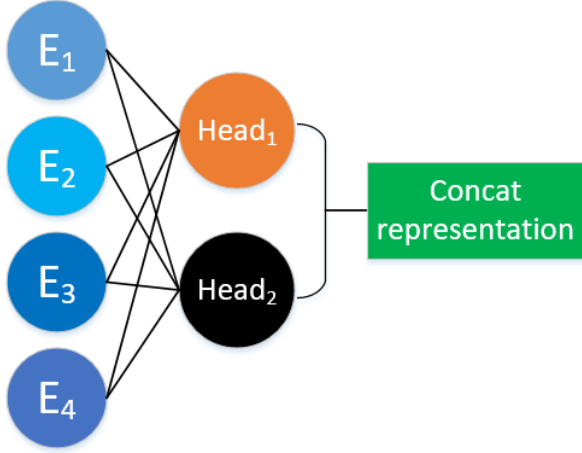


Figure 2. An example of two-heads events aggregation

### 3.2. Attentional Events Aggregation

It is difficult to extract the information through the long sequential data for the two reasons: (1) Within the long sequential record, the interaction among heterogeneous events could be complex, it is difficult to capture the dynamic interaction events representation; (2) The total dimensional numbers of the heterogeneous events embedding could be disastrously vast, making it impossible for the sequential model to capture the temporal representation.

To effectively capture the dynamic heterogeneous events representation, we propose attentional multihead aggregation. Given a time step events embedding  $E_t \in \mathbb{R}^{40 \times d}$ ,  $M$  randomly initialized mask-vectors are generated. Different masks are used to capture different aspects of events-interaction at  $time_t$  with attention-based mechanism. At last, all heads are concatenated to produce the eventual aggregation representation. The details of the aggregation are shown as follows:

$$\mathbf{h}_i = \sum_{j=1}^{40} attscore_i(K_t[j], \mathbf{m}_i)(W_{vi}E_t[j]) \quad (5)$$

$$attscore_i(\mathbf{k}, \mathbf{m}) = Softmax((W_{ki}\mathbf{k}) \cdot \mathbf{m}) \quad (6)$$

$$\mathbf{a} = Concat(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m) \quad (7)$$

Where  $\mathbf{h}_i \in \mathbb{R}^d$  is the  $i^{th}$  head of aggregation representation,  $\mathbf{a} \in \mathbb{R}^{M \cdot d}$  concatenates all heads,  $attscore_i$  is the attention mechanism of the  $i^{th}$  Head, getting the aggregation proportion of each event with calculating the dot product between events and Mask. Each head could capture its own concerned events information with its corresponding formed transform matrices  $W_k$ ,  $W_v$  and Mask vector  $\mathbf{m}$ . An example of two-heads aggregation is shown as Figure 2.

### 3.3. Sequential Model and Prediction

For each time step, we get an aggregation representation. Given  $X \in \mathbb{R}^{L \times 40}$ , a temporal events aggregation representation  $\mathbf{a} \in \mathbb{R}^{L \times (M \cdot d)}$  is captured. As LSTM is successfully used in sequential data, we pass on  $\mathbf{a}$  to one-layer bidirectional LSTM module. We sum up the last-time outputs of both forward and backward units and get the logits through a single dense layer with sigmoid activation. Our objective is a binary classification, we use cross entropy:

$$L(y, \hat{y}) = -(\hat{y} \cdot \log(y) + (1 - \hat{y}) \cdot \log(1 - y)) \quad (8)$$

## 4. Experiment Results

Table 1. local partitioned test metrics for different model

Model	AUC	APC	Utility score
MLP	0.7723	0.0711	0.2413
Transformer	0.8013	0.0923	0.3619
LSTM	0.8165	0.1040	0.3723
1-head-HEA-Transformer	0.8092	0.0955	0.3775
1-head-HEA-LSTM	0.8241	0.1120	0.3877
8-heads-HEA-Transformer	0.8264	0.1236	0.3968
8-heads-HEA-LSTM	0.8400	0.1307	0.4096
16-heads-HEA-LSTM	<b>0.8410</b>	<b>0.1314</b>	<b>0.4126</b>

### 4.1. Implementation Details

The provided data is divided into train and test sets at the ratio of 7 to 3. To conduct a further experiment, we divide the data into train and held sets at the ratio of 9 to 1, and then 5-fold cross validation is established. To evaluate our proposed module, we directly use a MLP prediction model, Transformer and LSTM with dense layer embedding, set various numbers of heads to gain different models. The results show that our proposed model, with high efficiency, can obviously improve the performance. We keep L as 24 (one day) in all experiments.

**Multihead aggregation:** We keep dimensional numbers(d) as 16 and set heads to be 1, 8 and 16 as 3 different aggregation modules.

We measure Area Under the Receiver Operating Characteristic curve (AUC) and Area Under The Precision Recall Curve (APC) as our evaluation metrics. What's more, the utility score function defined in CinC2019 challenge is used as the extra metric.

### 4.2. Baseline

**MLP:** Without considering the temporal information, Multi-layer perception can be used to directly model the raw record.

Table 2. local cross-validation metrics for HEA

Model	AUC	APC	Utility score(average/major vote/any vote)
8-heads-HEA-Transformer	0.8186	0.1049	0.3641/0.3635/0.3658
8-heads-HEA-LSTM	0.8206	0.1031	0.3656/0.3613/0.3643
16-heads-HEA-LSTM	<b>0.8224</b>	<b>0.1052</b>	0.3817/0.3750/ <b>0.3830</b>

**Dense Layer Embedding:** Given a one-time sampled record  $\mathbf{x}$ , the dense representation layer use a single dense layer with activation to generate the representation  $\hat{\mathbf{x}} \in \mathbb{R}$ .

**Transformer / LSTM:** Conventional sequential model Transformer and LSTM use a single dense layer embedding to capture the events representation.

### 4.3. Result

We firstly conduct an experiment for different models over train and test sets. The results over the test set are shown as Table 1. It should be noticed that all the metrics on both Table 1 and Table 2 are based on the locally partitioned test set from the public dataset.

The result shows that heterogeneous events aggregation modules could improve the metrics obviously, then we have the further experiment over 5-fold cross validation. The results over the held set are as Table 2 shows.

What’s more, our proposed model is in high efficiency, for it just needs to calculate the attention score between events and multiple heads. With a GeForce GTX 1080 10G, our proposed model with 16 heads cost 10 minutes per epoch to train over 10240000 training samples.

In thePhysioNet/Computing in Cardiology Challenge 2019, we got the results of score (0.402, 0.386, -0.169) on the test set A,B and C, with the overall utility score of 0.321, ranking the 13th out of 78 teams.

## 5. Conclusion

We proposed an attention-based sequential representation model to do early sepsis prediction from clinical data. Our proposed model includes two main parts: clinical events interaction extraction with heterogeneous events aggregation and temporal interaction capture with LSTM. Experiments in the PhysioNet/Computing in Cardiology Challenge 2019 show that the heterogeneous event aggregation module can shorten the length of clinical event sequence for better temporal dependency modeling, and the separated storage strategy of aggregation representation with different heads retains temporal interactions of events.

## Acknowledgments

This paper is partially supported by National Key Research and Development Program of China with Grant No. 2018AAA0101900, Beijing Municipal Commission of Science and Technology under Grant No. Z181100008918005, and the National Natural Science Foundation of China (NSFC Grant No. 61772039 and No. 91646202).

## References

- [1] Reyna M, Josef C, Jeter R, Shashikumar S, Westover M, Nemati S, Clifford G, Sharma A. Early prediction of sepsis from clinical data: the physionetcomputing in cardiology challenge 2019. *Critical Care Medicine* 2019 In press;.
- [2] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 2017;.
- [3] Qian F, Gong C, Liu L, Sha L, Zhang M. Topic medical concept embedding: Multi-sense representation learning for medical concept. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017; 404–409.
- [4] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation* 1997;9(8):1735–1780.
- [5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In *Advances in Neural Information Processing Systems*. 2017; 5998–6008.
- [6] Che Z, Purushotham S, Li G, Jiang B, Liu Y. Hierarchical deep generative models for multi-rate multivariate time series. In *International Conference on Machine Learning*. 2018; 783–792.
- [7] Liu L, Li H, Hu Z, Shi H, Wang Z, Tang J, Zhang M. Learning hierarchical representations of electronic health records for clinical outcome prediction. In *AMIA Annual Symposium*. 2019; .
- [8] Liu L, Shen J, Zhang M, Wang Z, Tang J. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018; .

Address for correspondence:

Ming Zhang  
1628, No.1 Science Building, Peking University, Beijing, China  
mzhang\_cs@pku.edu.cn