ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling

Yuan Zhang[1] , **Xi Yang**[1] , **Julie Ivy**[2] and **Min Chi**[1]
[1]Computer Science, North Carolina State University
[2]Industrial and System Engineering, North Carolina State University
[1]{yzhang93, yxi2, mchi}@ncsu.edu, [2]jsivy@ncsu.edu

Abstract

Modeling patient disease progression using Electronic Health Records (EHRs) is critical to assist clinical decision making. Long-Short Term Memory (LSTM) is an effective model to handle sequential data, such as EHRs, but it encounters two major limitations when applied to EHRs: it is unable to interpret the prediction results and it ignores the irregular time intervals between consecutive events. To tackle these limitations, we propose an attention-based time-aware LSTM Networks (ATTAIN), to improve the interpretability of LSTM and to identify the critical previous events for current diagnosis by modeling the inherent time irregularity. We validate ATTAIN on modeling the progression of an extremely challenging disease, septic shock, by using real-world EHRs. Our results demonstrate that the proposed framework outperforms the state-of-the-art models such as RETAIN and T-LSTM. Also, the generated interpretative time-aware attention weights shed some light on the progression behaviors of septic shock.

1 Introduction

Electronic Health Records (EHRs) are a large-scale and systematic collection of temporal health information of patients. The broad adoption of EHRs in medical systems has promoted the development of various computational methods for understanding the medical history of patients and predicting risks [Marlin *et al.*, 2012; Choi *et al.*, 2016a; Zhou *et al.*, 2013; Choi *et al.*, 2016b]. In this work, we focus on a task named *Disease Progression Modeling* (DPM), which monitors the disease developing process and predicts future risks based on patients' historical information. DPM is crucial for making clinical decisions and providing prompt medications. A large amount of recent works have been developed for this task [Choi *et al.*, 2016a; Esteban *et al.*, 2016; Lipton *et al.*, 2015; Zhou *et al.*, 2013]. Among them, Recurrent Neural Network (RNN) is one of the most extensively researched deep neural networks to handle the sequential data. As an extension of RNN, the Long-Short Term Memory (LSTM) is specifically designed to capture long-term patterns that commonly exist over a long period of patients' records [Sundermeyer *et al.*, 2012]. LSTM-based approaches have found a huge success in a variety of tasks involving sequential data, such as video processing, climate changes detecting, and EHRs representation learning [Jia *et al.*, 2019; Jia *et al.*, 2017; Lipton *et al.*, 2015; Esteban *et al.*, 2016].

Despite its great success, significant barriers remain when applying the standard LSTM for modeling EHRs. First, it cannot interpret prediction results. Second, it does not consider irregular time intervals between consecutive events.

Interpretability of computational models is extremely critical in healthcare-related domains. In real hospital settings, it is generally more important to learn about the discriminative *interpretable* patterns which capture informative progression of a disease than to induce an accurate predictive computational model. Various attention-based neural networks are widely developed to generate interpretations for EHRs. [Choi *et al.*, 2016b; Ma *et al.*, 2017; Sha and Wang, 2017]. As frontier work, RETAIN [Choi *et al.*, 2016b] applied a two-level attention mechanism to identify meaningful visits and specific features that contribute to the prediction.

Measurements in EHRs are commonly acquired with *irregular intervals*. For example, when a patient is in a severe condition, events are likely to be recorded more frequently than when a patient is in a relatively "healthier" condition. Hence such varying time intervals can reveal patient's health status on certain impending conditions, and it is important to consider the time intervals between temporal events to capture latent progressive patterns of a disease. There have been several previous works on handling the time irregularity [Baytas *et al.*, 2017; Pham *et al.*, 2016; Choi *et al.*, 2016a; Che *et al.*, 2017], e.g, Time-aware LSTM (T-LSTM) [Baytas *et al.*, 2017] transforms time intervals into weights and uses them to adjust the memory passed from previous moments.

In this work, we propose **ATTAIN**, an **at**tention-based **t**ime-aware dise**a**se progress**i**o**n** model, that incorporates the attention mechanism and models the time irregularity between events. Specifically, we adjust the memory of LSTM when accumulating previous information. Instead of adding memory from one previous event, we retrospect memories of all/several previous events and discount them by weights generated from attention mechanism and the time intervals between those events and current event. The overall weights represent how important each previous event is for the current event to identify the progressing condition. Three attention

mechanisms are explored: *global* ($\boldsymbol{g}$), *local* ($\boldsymbol{l}$), and *flexible* ($\boldsymbol{f}$), to generate the attention weights. On the other hand, the time intervals are transformed to decay weights through a decay function so that the outdated events are more likely to play a less important role than recent events for predicting the outcome of the current event. We believe the obtained attention weights together with the decay weights would not only lead to an effective predictive model but also result in an interpretable and clinically reasonable model.

We validate the proposed ATTAIN on the task of early prediction of septic shock. Sepsis is a life-threatening organ dysfunction [Singer *et al.*, 2016] and a leading cause of death in the United States. Septic shock, the most severe complication of sepsis, leads to a mortality rate as high as 50% and an increasing annualized incidence [Dellinger *et al.*, 2008]. In fact, as many as 80% of sepsis deaths could be prevented with timely diagnosis and treatment [Kumar *et al.*, 2006]. One major challenge associated with such early prediction is the subtle but fast progression at early stage. For example, only minor changes are reflected on white blood cells and body temperature at early stage [Kumar *et al.*, 2006]. Besides, the indicators of sepsis are non-specific, such as infection or fast heart rate. Hence, patients with such symptoms are highly likely to progress to other disease. Because of such delicate progressions, variables in the before-shock stage may either be measured infrequently or not measured at all. Therefore, it is quite critical and challenging to identify previous indicative moments and give accurate predictions of patients' health status.

Our experimental results on real-world EHRs show that ATTAIN outperforms the state-of-the-art models such as RETAIN and T-LSTM . Also, the generated interpretative time-aware attention weights shed some light on the progression behaviors of septic shock.

2 Method

2.1 Problem Definition

Our dataset can be represented as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, where N is the total number of hospital visits. It is composed of multi-variate irregular time series data and each visit $\boldsymbol{x}_k$ consists of a sequence of events: $\boldsymbol{x}_k = \{\boldsymbol{x}_k^1, ..., \boldsymbol{x}_k^{T_k}\}$, where $\boldsymbol{x}_k^t$ represents the patient's records at time step t in $\boldsymbol{x}_k$. We have $\boldsymbol{x}_k^t \in \mathbb{R}^D$, where D is the number of features or measurements recorded at each event and T_k is number of events in the visit k which varies with different visits.

For each $\boldsymbol{x}_k$, we are provided with the event-level label $\boldsymbol{y}_k = \{y_k^1, ..., y_k^{T_k}\}$ for the sequence of events. $y_k^t = 1$ indicates a patient is in septic shock state at a given time t in the visit x_k, otherwise $y_k^t = 0$. The goal of this work is to predict the $(t+1)$-th event-level label y_k^{t+1} given the clinical events from time 1 to t: $\boldsymbol{x}_k^1, \boldsymbol{x}_k^2, ..., \boldsymbol{x}_k^t$ in each visit k. For simplicity, we omit index k hereinafter when it does not cause ambiguity.

2.2 Long Short-Term Memory (LSTM)

In the standard LSTM cell unit, the cell state $\boldsymbol{c}^t$ serves as an internal memory and controls the information flow. It is generated by forgetting information through a forget gate $\boldsymbol{f}^t$ and most recent cell state $\boldsymbol{c}^{t-1}$, and adding new information through an input gate $\boldsymbol{i}^t$ and a candidate cell state $\tilde{\boldsymbol{c}}^t$ as listed in Eq. 1.

$$\begin{aligned} \boldsymbol{i}^t &= \text{sigmoid}(\boldsymbol{W}_h^i \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^i \boldsymbol{x}^t), \\ \tilde{\boldsymbol{c}}^t &= \tanh(\boldsymbol{W}_h^c \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^c \boldsymbol{x}^t), \\ \boldsymbol{f}^t &= \text{sigmoid}(\boldsymbol{W}_h^f \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^f \boldsymbol{x}^t), \end{aligned} \tag{1}$$

where $\boldsymbol{h}^{t-1}$ is a hidden state output by $\boldsymbol{c}^{t-1}$, $\{\boldsymbol{W}_h \in \mathbb{R}^{H\times H}, \boldsymbol{W}_x \in \mathbb{R}^{H\times D}\}$ denote network parameters to be trained and H is the number of hidden nodes. The new cell state can be obtained as follows:

$$\boldsymbol{c}^t = \boldsymbol{f}^t \otimes \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \otimes \tilde{\boldsymbol{c}}^t, \tag{2}$$

where $\otimes$ denotes entry-wise product.

Finally, we generate the hidden states by filtering the new cell state through an output gate layer $\boldsymbol{o}^t$, and produce the probability of each event t at risk of septic shock using a sigmoid function with parameter $\boldsymbol{U}$:

$$\begin{aligned} \boldsymbol{o}^t &= \text{sigmoid}(\boldsymbol{W}_h^o \boldsymbol{h}^{t-1} + \boldsymbol{W}_x^o \boldsymbol{x}^t), \\ \boldsymbol{h}^t &= \boldsymbol{o}^t \otimes \tanh(\boldsymbol{c}^t), \\ \boldsymbol{p}^t &= \text{sigmoid}(\boldsymbol{U}\boldsymbol{h}^t). \end{aligned} \tag{3}$$

2.3 ATTAIN Networks

Fig. 1 shows the architecture of proposed ATTAIN framework. In the original LSTM, the cell state of current event $\boldsymbol{c}^t$ obtains memory from its most recent event i.e., $\boldsymbol{c}^{t-1}$ as shown in Eq. 2. Due to the well-known vanishing gradient problem [Baytas *et al.*, 2017], the learning of each cell state still heavily depends on most recent events. However, in disease progression, relationships of events are non-trivial in that the current condition is often not impacted only by recent moments. Single memory-reading also brings that it cannot interpret how critical of each past event. Conversely in practice, doctors usually review the patient's records to identify the critical previous events.

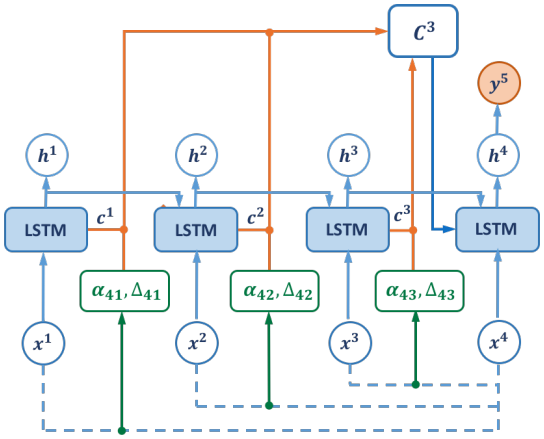

Figure 1: Illustration of proposed ATTAIN model. ATTAIN accumulates the cell states from all the previous events and regularize the old memories through two sources of weights generated from attention mechanism ($\boldsymbol{\alpha}$) and time intervals ($\boldsymbol{\Delta}$).

Here we propose to use attention mechanism to collect multiple previous memories and explore the relationships of events in patient trajectories. We expect it can help doctors make prompt decisions and achieve a better prediction.

The standard LSTM implicitly assumes that intervals between time steps, or events, in a sequence are uniformly distributed, whereas the frequency of collecting patients' records often vary greatly based on many factors. The interval between two consecutive events during a patient's visit can span from minutes to several weeks. In general, the events that occur long time ago tend to have less impact to the current event and thus we should properly reduce their contributions. Therefore, it is important to consider the elapsed time when predicting the current event's output. In addition, these irregular time intervals might be indicative for patients' health status on certain impending conditions. For example, when a patient is in a more severe condition, events are likely to be recorded more frequently than when a patient is in a relatively "healthier" condition.

To incorporate the elapsed time into the standard LSTM, we feed the intervals into a monotonic decreasing function. In specific, for each event, the memories collected from its previous events are discounted by a decay factor, which is estimated based on the time intervals between previous events and the current event. In this way, if the previous event occurs a long time ago, the dependency from the previous moment would not play an active role for predicting the current event's output.

As shown in Fig. 1, in the proposed ATTAIN, instead of reading the information from just one previous cell state, we collect all/some previous cell states and regularize these old memories by weights generated by attention mechanism and time intervals as follows:

$$\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{t-1} \alpha_{ti} \cdot \boldsymbol{c}^i \cdot g(\Delta t_{ti}), \tag{4}$$

where α_{ti} is the attention weight from i-th event to the current event t, Δt_{ti} is the time interval between i-th event to the current event, $g(\cdot)$ is a decay function, and m stands for the number of events to look backwards. Then we replace $\boldsymbol{c}^{t-1}$ in Eq. 2 with $\boldsymbol{C}^{t-1}$ for generating the new cell state.

Three Attention Mechanisms

To determining m, three ways are explored:

- *Global Attention* generates attention weights α_{ti} from *all* the previous events for the current time step t, , i.e, $m = t-1$ in Eq. 4.

$$\begin{aligned} e_{ti} &= (\boldsymbol{x}^i)^\top \boldsymbol{W}_\alpha \boldsymbol{x}^t, \text{for } i = 1, ..., t-1 \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t1}, e_{t2}, ..., e_{t(t-1)}) \end{aligned} \tag{5}$$

  The advantage of the global attention is that it considers all past events and its disadvantage is that it is computationally expensive and attention weights are sparsely distributed. To address these problems, we also explore the following two options.

- *Local Attention* only considers a *fixed* number of previous events, m in Eq. 4. The hyperparameter m can be optimized by grid search or suggested by clinicians.

$$\begin{aligned} e_{ti} &= (\boldsymbol{x}^i)^\top \boldsymbol{W}_\alpha \boldsymbol{x}^t, \text{for } i = t-m, ..., t-1 \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t(t-m)}, e_{t(t-m+1)}, ..., e_{t(t-1)}) \end{aligned} \tag{6}$$

- *Flexible Attention* learns the optimal number of previous events adaptively for different given events, i.e, m varies for different time step t. The intuition is that different number of previous events should be considered for different given events.

$$\begin{aligned} m(t) &= \lceil (t-1) \cdot \text{sigmoid}(\boldsymbol{v}_m^\top \tanh(\boldsymbol{W}_m \boldsymbol{x}^t)) \rceil \\ \boldsymbol{\alpha}_t &= \text{softmax}(e_{t(t-m(t)}, e_{t(t-m(t)+1)}, ..., e_{t(t-1)}), \end{aligned} \tag{7}$$

  where $\boldsymbol{W}_m$ and $\boldsymbol{v}_m$ are parameters to be learned, and $m(t) \in [1, t-1]$.

Previously RETAIN [Choi *et al.*, 2016b] introduced attention mechanism in learning health record representation. RETAIN generated attention weights from the hidden states of RNN: $e_{ti} = \boldsymbol{W}_\alpha^\top \boldsymbol{h}^i, i \in [1, t-1]$. However, it does not capture the relation between the targeting event t and past events is. Different with RETAIN, the aforementioned three attention mechanisms incorporate this relation when estimating attention weights. Additionally, when aggregating previous information (see Eq. 4), α_{ti} measures the weight of event i for the current event t and hence its computation should solely depend on the data at events i and t. Therefore we use the original inputs $\boldsymbol{x}$ (i.e. $\boldsymbol{x}^t$ and $\boldsymbol{x}^i$) instead of hidden states $\boldsymbol{h}$ (i.e. $\boldsymbol{h}^t$ and $\boldsymbol{h}^i$) since the pair $[\boldsymbol{x}_t, \boldsymbol{x}_i]$ extracts immediate relations among the medical variables of two events while $[\boldsymbol{h}_t, \boldsymbol{h}_i]$ encodes the information from multiple events.

Time-aware Decay Function

Time intervals might range from minutes, hours and even to days in healthcare domain. We integrate the intervals into the model and make LSTM predict in a time-aware way. As a guideline, the function in Eq. 8 is validated and suggested by several prior works for long time periods in EHR data [Pham *et al.*, 2016; Baytas *et al.*, 2017].

$$g(\Delta t) = 1/\log(e + \Delta t) \tag{8}$$

Previously T-LSTM [Baytas *et al.*, 2017] also integrated the elapsed time to adjust the cell memory. It divided the previous cell memory as short-term ($\boldsymbol{c}_S^{t-1}$) and long-term ($\boldsymbol{c}_L^{t-1}$), then discounted only the short-term memory through the same decay function, i.e., $\boldsymbol{c}^t = \boldsymbol{f}^t \otimes (\boldsymbol{c}_L^{t-1} + \boldsymbol{c}_S^{t-1} \cdot g(\Delta t)) + \boldsymbol{i}^t \otimes \tilde{\boldsymbol{c}}^t$. T-LSTM depends on only one previous cell state to adjust the memory and is unable to interpret how important each event is for predicting the current outcome.

In summary, through the aforementioned memory adjustment, ATTAIN would identify the critical previous moments and take the effect of elapsed time into account. The two sources of weights, based on attention mechanism and time intervals, are simultaneously tuned together. The intuitions behind ATTAIN can be seen as mimicking doctors reading records for diagnosing in the real world in that they generally pay attention to both very important indicative previous events even though they happened long time ago but also the most recent events. As a result, we expect ATTAIN can lead to more reasonable predictions than the standard LSTM.

3 Experiments

3.1 Data Description

Our EHR data was collected from Christiana Care Health System Health System (CCHS) from July, 2013 to December, 2015. Each data point is a visit with a series of events(see 2.1). In total, there are 210,289 visits, and 10,412,729 medical events. The study population is patients with *suspected infection*, which is identified by the presence of any type of antibiotic, antiviral, or antifungal administration, or a positive test result of Point of Care Rapid, and it consists of 52,919 visits and 4,224,567 medical events. The definition of study population and the following data prepossessing were determined by leading clinicians from CCHS and Mayo Clinic.

After preprocessing, our final dataset contains 2,100 *visits* (1,869 positives and 231 negatives) and 209,346 *events* (22,430 positives and 186,916 negatives).

Tagging. International Classification of Diseases, Ninth Revision (ICD-9) are widely used for clinical labeling (i.e., septic shock or not). However, as visit-level labels, ICD-9 codes cannot tell when septic shock occurs at event level, and they are not persistently reliable [Ho *et al.*, 2014]. Therefore based on Third International Consensus Definitions [Singer *et al.*, 2016], our clinicians identified septic shock *at each event* as having received vasopressor(s) or persistent hypotension (systolic blood pressure<90 mmHg or mean arterial pressure<65 mmHg for more than 1 hour). When applying both ICD-9 and our tagging rule, we identify 1,869 shock positive visits and 23,901 negative ones.

Sampling. We further conduct stratified random sampling on negative visits while keeping their underlying distribution of age, gender, race and length of stay the same as positive visits, and then get 1,869 negative visits. Given the number of negative events are dominant in both positive and negative visits, we randomly sample 231 out of 1,869 negative visits to maintain the ratio of positive events around 15%.

Data selection. We focus on the visits in range of 3-90 days since short visits do not hold sufficient information for analysis and long visits introduce data sparsity. Importantly, we exclude the visits that developed septic shock within 8 hours after admission, because our clinicians suggest that such patients generally already show septic-related symptoms on admission and doctors can easily catch such cases.

Feature selection. We exclude features with a missing rate more than 90% and our features can be divided into four categories: 1) vital signs: heart rate, temperature, etc; 2) lab results: BUN, creatinine, white blood cell count (WBC), 18 culture tests, etc; 3) interventions: FIO2, oxygen flow, etc; 4) locations (e.g, emergency or nurse), descriptions, identifiers.

Data aggregation and imputation. We aggregate the events within an hour to an event and take the mean measurement within 1-hour window. For frequently measured features (e.g. vital signs), we add statistical features (e.g. min and max) measured in the 1-hour window. Suggested by our clinicians, we impute missing values with the last value within the fixed length of forward time window (8 hours for vital signs and 24 hours for lab tests) and impute the remaining ones with feature-wise mean values.

3.2 Experimental Setup

Baseline Approaches

We compare ATTAIN with three baselines: 1) LSTM without either attention or time-aware mechanisms; 2) RETAIN [Choi *et al.*, 2016b] with a two-level attention mechanism at both the event-level and the variable level; 3) T-LSTM [Baytas *et al.*, 2017] with time-aware mechanism only.

Our Approaches

We integrate three attention mechanisms with the standard LSTM: LSTM$_g$, LSTM$_l$, LSTM$_f$. The attention weights are generated by Eq. 5 - 7 respectively. Old memory is obtained by accumulating all or selective previous cell states:

$$\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{t-1} \alpha_{ti} \cdot \boldsymbol{c}^{i}. \tag{9}$$

Specifically for local attention, we find the optimal setting of m by grid search on a validation set (described later). Fig. 2(a) shows the model achieves competitive performance using tenth-interval $m \in \{20, 30, 40\}$. We hence further explore the performance for $m \in [20, 40]$ incremented by 1 and Fig. 2(b) shows that the optimal values are $m \in \{22, 26\}$. Here we use $m = 24$.

Furthermore, we add the time-aware mechanism to the three models above resulted in three ATTAIN models named ATTAIN$_g$, ATTAIN$_l$, ATTAIN$_f$ respectively. The $\boldsymbol{C}^{t-1}$ is calculated based on Eq. 4 in a way that the old memory is regularized by both attention weights and decay weights.

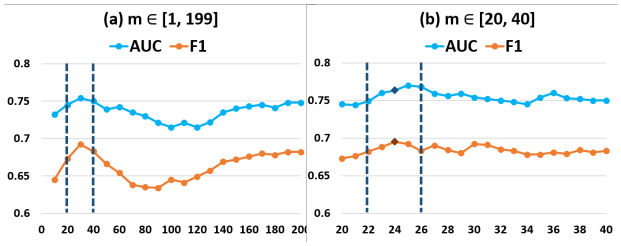

Figure 2: AUC and F1 score of LSTM$_l$ when: (a) m is taken from $[1, 199]$ with interval 10. (b) $m \in [20, 40]$.

Evaluation and Setting

Our evaluation metrics include sensitivity/recall, specificity, precision/positive predictive value (PPV), F1-score, and area under the ROC (receiver operator characteristic) curve (AUC) [Nachimuthu and Haug, 2012]. Precision, recall, F1 and AUC are widely used for evaluating machine learning approaches. In healthcare domain, researchers commonly refer to sensitivity, specificity and PPV for the annotation performance.

In our implementation, all the models are updated using mini-batch stochastic optimizer and the batch size is 50. The training epochs is 50 with early stopping, the learning rate is 0.01, and the number of hidden units for LSTM is 72. In training process, we randomly divide the data sets into the training, validation and testing set with the ratio of 70%, 15%, and 15%. Each experiment is repeated 10 times with random model initialization and we report the average values with standard deviation for each evaluation metric.

Method	A	T	Sensitivity/Recall	Specificity	PPV/Precision	F1-score	AUC
LSTM	-	-	0.627(±0.023)	0.632(±0.021)	0.635(±0.020)	0.631(±0.021)	0.716(±0.020)
RETAIN	✓	-	0.618(±0.015)	0.654(±0.016)	0.651(±0.016)	0.634(±0.016)	0.732(±0.010)
T-LSTM	-	✓	**0.643**(±0.009)	**0.680**(±0.012)	**0.702**(±0.013)	**0.671**(±0.010)	**0.745**(±0.013)
LSTM_g	✓	-	0.628(±0.013)	**0.798**(±0.015)	**0.747**(±0.018)	0.682(±0.016)	0.748(±0.014)
LSTM_l	✓	-	**0.684**(±0.007)	0.742(±0.013)	0.707(±0.012)	**0.695**(±0.011)	**0.763**(±0.008)
LSTM_f	✓	-	0.667(±0.022)	0.731(±0.017)	0.726(±0.016)	0.695(±0.019)	0.755(±0.016)
ATTAIN_g	✓	✓	0.636(±0.016)	⋆ **0.818**(±0.008)	⋆ **0.803**(±0.010)	0.710(±0.014)	0.782(±0.015)
ATTAIN_l	✓	✓	⋆ **0.695**(±0.014)	0.746(±0.012)	0.744(±0.015)	0.718(±0.014)	0.804(±0.010)
ATTAIN_f	✓	✓	0.686(±0.018)	0.767(±0.016)	0.758(±0.016)	⋆ **0.720**(±0.017)	⋆ **0.811**(±0.011)

· The best values of each metric within the session are in bold, and the best values of each metric across all the models are labeled with ⋆.

Table 1: Performance (±standard deviation) of baselines and our approaches on septic shock overall prediction.

4 Results

4.1 Results of Overall Prediction

Table 1 shows the performance of all the models and from top to bottom, we have three baseline models, our proposed three attention mechanisms with LSTM, and the three corresponding ATTAIN models. Columns 'A' and 'T' indicate whether attention or time-aware mechanisms are used in the models. For each sub-session, the best results are marked in bold and the best model across all methods are labeled with ⋆.

Among the three baselines, both RETAIN and T-LSTM outperform the standard LSTM on almost all the evaluations, which shows that either the attention mechanism or time-aware mechanism can lead to a better prediction of the disease progression. As in the original work [Choi *et al.*, 2016b], RETAIN does not boost from LSTM much and this might come from two reasons. First, their attention mechanism did not incorporate the relation between the previous events and the current event. Second, due to high missing rate of features, the variable-level attention did not take into full effect.

All three attention-based models (in the middle sub-session) outperform three baselines across all evaluations with only one exception that LSTM_g has worse recall than T-LSTM. When comparing the three attention mechanisms, LSTM_l performs more stably than LSTM_g and LSTM_f, i.e., 0.695 vs. 0.682 and 0.695 (larger std) for F1 score and 0.763 vs. 0.748 and 0.755 for AUC. We found that the attention weights of LSTM_g are sparsely distributed in long sequences and cannot effectively detect critical moments. The result of LSTM_l validates that $m = 24$ is clinically reasonable. When only using attention mechanism, though close, LSTM_f, deciding how many previous events to pay attention in a data-driven manner, is not as competitive as LSTM_l.

Except that ATTAIN_g has a worse recall than T-LSTM, all three ATTAIN models (in the bottom sub-session) outperform three baselines with obvious improvement. Moreover, when comparing between ATTAIN and our three attention-based models, ATTAIN_g outperforms LSTM_g, and the same observation on ATTAIN_l over LSTM_l and ATTAIN_f over LSTM_f . Thus, when integrating both attention and time-aware mechanisms , the two sources of weights allow the model to comprehensively identify the meaningful previous events in terms of medical relation and time effect. For example, if attention weights indicate two previous events are equally important, we pay attention to the recent event first, exactly as doctors perform in real world.

Overall, ATTAIN_f achieves the best F1-score and AUC. ATTAIN_l is close to ATTAIN_f, i.e. 0.718 vs. 0.720 on F1 and 0.804 vs. 0.811 on AUC and gets the best recall. Since our data are aggregated at hourly rate, 24-event retrospect guarantees attending the last 24 hours of a patient status. In practice the clinicians prefer this local-based model because previous information can be sufficiently obtained with certainty.

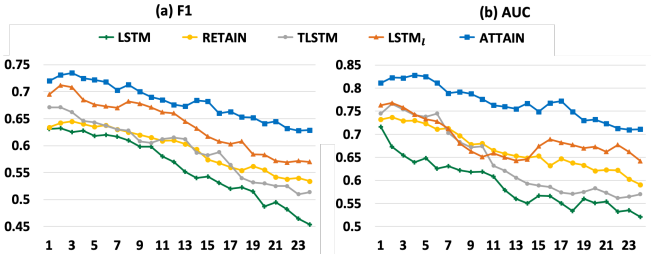

Figure 3: (a) F1-score of early prediction at different hours. (b) AUC of early prediction at different hours.

4.2 Results of Early Prediction

Our best attention-based model LSTM_l and our best approach ATTAIN_f are compared against three baselines: LSTM, RETAIN and T-LSTM for early prediction, i.e. to detect the patients who are developing septic shock in advance. To do so, we utilize the future diagnostic condition as labels for each current event in training process. For example, in previous setting, we use $\{\boldsymbol{x}_k^i\}_{i=1}^t$ to predict the $(t+1)$-th event-level label y_k^{t+1}, one event in advance. In early prediction setting, we will use $\{\boldsymbol{x}_k^i\}_{i=1}^t$ to predict $y^{t+\eta}$, η events in advance. Since our data is aggregated on an hourly basis, η-event early guarantees at least η-hour ahead. We track the performance of each model on F1-score and AUC from one-event earlier up to 24 incremented by 1.

Fig. 3 shows the results of early prediction on five models. As expected, as η increases, the early prediction tasks become more and more challenging in that the performance of all the models has gradually degraded. Despite this, ATTAIN_f has been decreasing much slower and steadily staying on the top of all other models in terms of both AUC and F1. LSTM_l follows ATTAIN_f to be the second best model and the standard LSTM performs the worst.

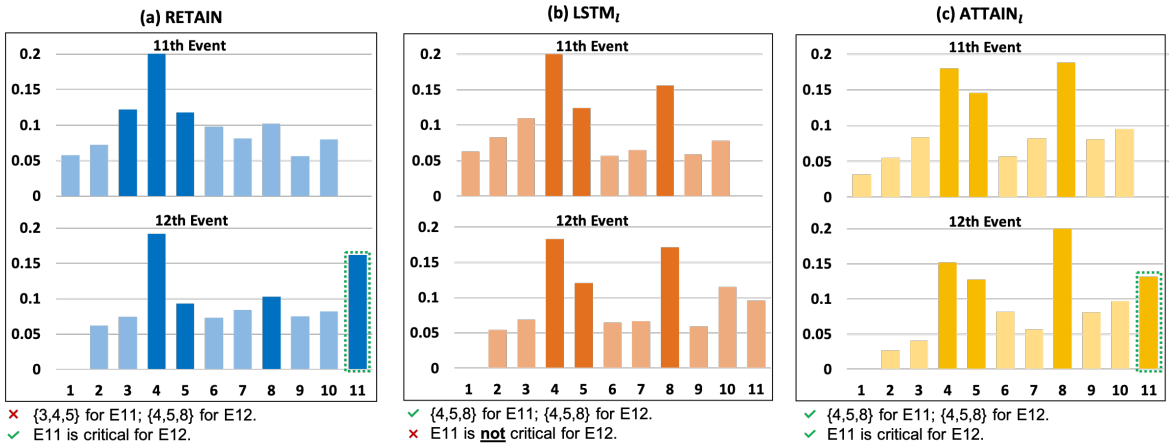

Figure 4: Attention weights for the 11th and 12th events achieved from (a) RETAIN; (b) LSTM$_l$; (c) ATTAIN$_l$.

4.3 Case Study

To illustrate the interpretability, Fig. 4 shows a case study by comparing the weights learned from RETAIN, LSTM$_l$ and ATTAIN$_l$ for a sample patient's visit. To simplify, we fix $m = 10$ for all three models and the weights are normalized. The patient is in septic shock at events 11 and 12. For event 11, we expect the models can identify some critical previous event(s). We consider *three* prior events with the highest weights to be *critical events*. At event 12, septic shock continues, we expect a reliable model: a) identifies event 11 to be critical for event 12, because the former clearly explains why the patient was still in septic shock at event 12, and b) identifies the same critical events (other than event 11) as those for event 11 for event 12. Fig. 4 shows that RETAIN indeed identifies that event 11 is critical for event 12, but it discovers different sets of critical events for the two events. LSTM$_l$ consistently detects the same set of critical events, $\{4, 5, 8\}$, for both events 11 and 12, but it fails to identify that event 11 is critical to event 12. Finally, our ATTAIN model successfully identify that event 11 is important and besides event 11, event 12 has the same critical set: $\{4, 5, 8\}$ as event 11.

5 Related Work

EHRs have been a popular research platform with increasing availability to develop predictive models for tasks of disease progression [Zhang *et al.*, 2017; Wang *et al.*, 2014], phenotyping [Che *et al.*, 2015; Baytas *et al.*, 2017; Liu *et al.*, 2015; Li *et al.*, 2015], diagnosis prediction [Choi *et al.*, 2016b], etc. However, EHRs also pose numerous challenges, such as they are noisy, fragmental and high dimensional. To this end, deep learning networks, such as RNN [Che *et al.*, 2018] and LSTM [Lin *et al.*, 2018; Yang *et al.*, 2018] can assist in learning complex relationships among medical events.

In recent years, attention mechanisms are extensively explored to interpret the model output and greatly improve the prediction performance. For example, RETAIN applies a reverse time attention mechanism in an RNN [Choi *et al.*, 2016b] and Dipole [Ma *et al.*, 2017] uses the similar attention networks for diagnosis prediction. Another challenge associated with EHR data, time irregularity, has also been tackled. T-LSTM [Baytas *et al.*, 2017] divides short-term from the previous cell memory, and adjusts it with a time-aware mechanism. In [Pham *et al.*, 2016], the time intervals are used to modify the forget gate of LSTM. In [Che *et al.*, 2018], time gaps are made regular through data imputation methods. Finally, Health-ATM [Ma *et al.*, 2018] extracts patient information patterns with attentive and time-aware models through RNN and Convolutional Neural Networks (CNN). Compared with the prior works, our proposed method explores different attention mechanisms to generate weights for the past events while handling the time irregularity in EHRs. For acute medical conditions such as septic shock, it is extremely significant to identify critical and timely meaningful events.

6 Conclusion

Disease progression modeling is an important task especially for acute medical conditions such as septic shock. In this work, we propose ATTAIN, an attention-based time-aware LSTM networks to effectively improve the interpretability of LSTM while also modeling the irregular time intervals. In specific, ATTAIN employs attention weights combining with time decay function to identify the contributions of the historical events to the current event. The experimental results on a real EHR dataset demonstrate the effectiveness of ATTAIN compared with state-of-the-art baselines. One limitation of this work is not generating feature-level attentions at each event, since the feature-level attentions can vary drastically over multiple events in acute disease conditions, especially given the impact of high missing rate in real-world EHRs. We keep this challenging task as our future work. We will be working on imputing realistic missing data for heterogeneous patient groups through domain adaptation and generative networks. Also, we will validate ATTAIN on other EHRs such as MIMIC-III and for other disease prediction tasks.

Acknowledgements

This research is supported by the NSF Grants: #1522107, #1651909, #1660878 and #1726550.

References

[Baytas *et al.*, 2017] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Paient subtyping via time-aware lstm networks. In *SIGKDD*. ACM, 2017.

[Che *et al.*, 2015] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *SIGKDD*. ACM, 2015.

[Che *et al.*, 2017] Chao Che, Cao Xiao, Jian Liang, Bo Jin, Jiayu Zho, and Fei Wang. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *SDM*. SIAM, 2017.

[Che *et al.*, 2018] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *MLHC*, pages 301–318, 2016.

[Choi *et al.*, 2016b] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, pages 3504–3512, 2016.

[Dellinger *et al.*, 2008] R Phillip Dellinger, Mitchell M Levy, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive care medicine*, 2008.

[Esteban *et al.*, 2016] Cristóbal Esteban, Oliver Staeck, et al. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *ICHI*, pages 93–101. IEEE, 2016.

[Ho *et al.*, 2014] Joyce Ho, Cheng Lee, and Joydeep Ghosh. Septic shock prediction for patients with missing data. *Management Information Systems*, 5(1), April 2014.

[Jia *et al.*, 2017] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Incremental dual-memory lstm in land cover prediction. In *SIGKDD*. ACM, 2017.

[Jia *et al.*, 2019] Xiaowei Jia, Sheng Li, Ankush Khandelwal, Guruprasad Nayak, Anuj Karpatne, and Vipin Kumar. Spatial context-aware networks for mining temporal discriminative period in land cover detection. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 513–521. SIAM, 2019.

[Kumar *et al.*, 2006] Anand Kumar, Daniel Roberts, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 2006.

[Li *et al.*, 2015] Hui Li, Xiaoyi Li, Xiaowei Jia, Murali Ramanathan, and Aidong Zhang. Bone disease prediction and phenotype discovery using feature representation over electronic health records. In *ACM-BCB*. ACM, 2015.

[Lin *et al.*, 2018] Chen Lin, Yuan Zhang, Min Chi, et al. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm. In *ICHI*, pages 219–228. IEEE, 2018.

[Lipton *et al.*, 2015] Zachary C Lipton, David C Kale, et al. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[Liu *et al.*, 2015] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *SIGKDD*, pages 705–714. ACM, 2015.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *SIGKDD*, pages 1903–1911. ACM, 2017.

[Ma *et al.*, 2018] Tengfei Ma, Cao Xiao, and Fei Wang. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *SDM*, pages 261–269. SIAM, 2018.

[Marlin *et al.*, 2012] Benjamin M Marlin, David C Kale, Robinder G Khemani, et al. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *IHI*, pages 389–398. ACM, 2012.

[Nachimuthu and Haug, 2012] Senthil K Nachimuthu and Peter J Haug. Early detection of sepsis in the emergency department using dynamic bayesian networks. In *AMIA*, 2012.

[Pham *et al.*, 2016] Trang Pham, Truyen Tran, Dinh Phung, et al. Deepcare: A deep dynamic memory model for predictive medicine. In *PAKDD*. Springer, 2016.

[Sha and Wang, 2017] Ying Sha and May D Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *BCB*, 2017.

[Singer *et al.*, 2016] Mervyn Singer, Clifford S Deutschman, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[Sundermeyer *et al.*, 2012] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *ISCA*, 2012.

[Wang *et al.*, 2014] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *SIGKDD*, pages 85–94. ACM, 2014.

[Yang *et al.*, 2018] Xi Yang, Yuan Zhang, and Min Chi. Time-aware subgroup matrix decomposition: Imputing missing data using forecasting events. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[Zhang *et al.*, 2017] Yuan Zhang, Chen Lin, Min Chi, et al. Lstm for septic shock: Adding unreliable labels to reliable predictions. In *Big Data*, pages 1233–1242. IEEE, 2017.

[Zhou *et al.*, 2013] Jiayu Zhou, Jimeng Sun, Yashu Liu, Jianying Hu, and Jieping Ye. Patient risk prediction model via top-k stability selection. In *SDM*, pages 55–63. SIAM, 2013.