LSTM for Septic Shock: Adding Unreliable Labels to Reliable Predictions

Yuan Zhang*, Chen Lin*, Min Chi*, Julie Ivy†, Muge Capan‡ and Jeanne M. Huddleston§
*Computer Science, North Carolina State University
†Industrial and System Engineering, North Carolina State University
‡Christiana Care Health System
§Mayo Clinic
*{yzhang93, clin12, mchi}@ncsu.edu, †jsivy@ncsu.edu
‡muge.capan@christianacare.org,§huddleston.jeanne@mayo.edu

***Abstract*—Sepsis is a leading cause of death over the world and septic shock, the most severe complication of sepsis, reaches a mortality rate as high as 50%. Early diagnosis and treatment can prevent most morbidity and mortality. Nowadays, the increasing availability of the electronic health records (EHRs) has generated great interests in developing models to predict acute medical conditions such as septic shock. However, septic shock prediction faces two major challenges : 1) how to capture the informative progression of septic shock in a long visit to hospital of a patient; and 2) how to obtain reliable predictions without well-established moment-by-moment ground-truth labels for septic shock.**

In this work, we proposed a generic framework to predict septic shock based on Long-Short Term Memory (LSTM) model, which is capable of memorizing temporal dependencies over a long period. The framework integrates two levels of imperfect yet informative labels to jointly learn the discriminative patterns of septic shock: ICD-9 code as the visit-level label and the clinical criteria designed by domain experts as the moment-by-moment event-level label. We evaluate our method on a real-world data extracted from an EHR system constituted by 12,954 visits and 1,348,625 events, and compare it against multiple baselines. The robustness of the method is validated using three sets of clinician-proposed adjusted ground-truth labels. Also, we explore whether the framework is effective for the early prediction of the patients developing septic shock. The experimental results demonstrate the superiority of our proposed method in the task of septic shock prediction.

I. Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection [26]. As a leading cause of death in the United States, sepsis accounts for nearly $24 billion (6.2%) of total hospital costs in 2013 [30]. Each year 1.6 million people are diagnosed with sepsis, more than prostate cancer, breast cancer and AIDS combined [7] [23]. Septic shock, the most severe complication of sepsis, reaches a mortality rate as high as 50% [20] and the annualized incidence keeps rising [6].

Prior studies have demonstrated that early diagnosis and treatment of septic shock can decrease patients' mortality and length of stay significantly [18] [5] [13]. As many as 80% of sepsis deaths could be prevented with rapid diagnosis and treatment [18]. Conversely, every hour of delay in antibiotic treatment leads to an 8% increase in mortality for septic shock patients [18]. There exist two major challenges associated with sepsis/septic shock prediction: the first one is the subtle progression of sepsis/septic shock in a long period and the second one is the lack of well-established definition for labeling sepsis/septic shock.

As for the first challenge, the clinical signs and symptoms at the early stage of sepsis/septic shock are often subtle and non-specific. For example, only minor changes are reflected on white cell count and body temperature at the early stage of sepsis. Moreover, infection, a hallmark of sepsis, is highly likely to progress to other disease and hence not a specific symptom for sepsis. Therefore, it is critical to learn about the discriminative patterns of sepsis/septic shock and capture the informative progression during a patient's stay. It becomes especially challenging when the whole stay spans over a long time since it is difficult for most of the sequential models to memorize so many details and persistently connect previous information to the present without much loss.

To overcome this challenge, we model the septic shock progression using a variant of Long-Short Term Memory (LSTM) networks, which has shown extensive prospects in a variety of sequential labeling applications, such as climate changes, health-care records, traffic monitoring, etc [19] [17] [16] [8]. Compared with other sequential models like Recurrent Neural Networks (RNN), LSTM is capable of memorizing temporal dependencies over a long period [27]. On the other hand, the electronic health records (EHRs) is a popular research platform with increasing availability to develop predictive models for acute medical conditions such as septic shock. However, EHR data also poses numerous challenges, such as they are noisy, fragmental and high dimensional. To this end the incorporation of temporal dependencies can assist in mitigating the impact of noise and in learning complex relationships among features.

The second challenge stems from no gold-standard definition or criteria of labeling sepsis/septic shock at any given time point. Due to varying purposes and expertise levels on the disease, different decision-making systems have certain *biases* on labelling sepsis/septic shock. During

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

each hospital visit, a patient usually takes multiple tests and gets measurements at different time points. The assessment for the entire visit provides a *visit-level* label while the assessment for each time point provides a *event-level* label. We argue that though informative, both levels of labels are imperfect and they do not even agree for many cases.

At the visit level, septic shock patients are identified through the so-called International Classification of Diseases, Ninth Revision (ICD-9) code (785.52) for billing purpose. As billing codes, ICD-9 is only coded for limited number of complications and dramatically different diseases can often share the same billing code as long as they have the same cost. Hence, ICD-9 may report false positives of septic shock and cannot fully represent the real medical states of patients. Indeed it has been widely argued that ICD-9 codes cannot be used for establishing reliable gold standards for various clinical conditions [10] [15].

At the event level, septic shock is usually determined by clinical criteria which is generally set based on current vital signs and lab tests, such as "hypotension" or "ongoing vasopressor therapy". The event-level labels are often noisy since the attributes/features involved in the criteria are not always observable and updated at each time point. For example, white cell count are only tested every 24 hours. Generally in this situation the unavailable values are filled with the ones from previous moments, which may lead to out-of-state evaluation for patients' states. Moreover, the values of some attributes return to normal after patients receive a treatment, such as vasopressor administration, resulting in a negative event-level label even though the patient is still in septic shock. Finally and most importantly, there exists no well-established clinical criteria on whether a patient is in septic shock state at each time point.

To summarize, although visit-level labels and event-level labels allow us to grasp certain useful information of the disease, we cannot directly utilize them as ground truth of septic shock. Therefore, we propose a framework to learn the sequential patterns of septic shock by jointly leveraging both levels of imperfect yet informative labels. Initially, the data are labeled at both visit level (ICD-9) and at event level (criteria defined by clinicians). In the training process, the framework first carries out a standard supervised learning cycle by training a LSTM model with event-level labels. Then we apply the learned LSTM model on the training data to calculate the probabilities that a patient is in the shock state at any given time point (event level). With the event-level output, we check the consistency between the event-level and the visit-level information (ICD-9), and revise the training objectives based on conditions. Then the trained LSTM model is updated using the training data with revised labels. The procedure is iteratively conducted until the model converges and no changes happen for the labels of training data. By utilizing this iterative training process, we expect the proposed framework to explore latent true labels of septic shock from combining two levels of noisy labels. In this way, the proposed framework can make accurate predictions to reflect real disease conditions.

Machine learning techniques have achieved considerable success in predicting sepsis/septic shock by using visit-level or event-level labels *but not both*. At visit level, most prior work used ICD-9 code as ground truth and considered the whole visit as one sample. For example, Brause *et al.* [3] used neural networks to predict the critical states of sepsis; support vector machines (SVM) was proved to be effective at sepsis-related classification tasks [32] [28]. To truly grasp the temporal dependencies and capture the subtle progression of septic shock, it is important to predict at event level. Peelen *et al.* [24] developed a dynamic Bayesian network (DBN) to model the progression of organ failure, a severe state developing into septic shock. Generally speaking, compared to works done at visit level, relatively less work was done at the event level for predicting septic shock, due to the fact that there is no well-established criteria on labelling septic shock at any time point. Some work relied on the labels manually annotated by clinicians [21], but such labeling process is often too expensive to be feasible for large volume of data. Therefore, we propose to construct a general data-driven framework that leverages multiple sources of imperfect yet informative labels.

We validate the proposed method on a real-world data extracted from an electronic health records (EHRs) system and compared it with multiple strong baselines. In practice there are various ways to manually adjust the data labels to approach ground truth, so we test the robustness of the framework by comparing its prediction results against three sets of ground-truth labels. Finally, we demonstrate that the proposed framework is also effective for early prediction of septic shock.

II. Problem Definition and Datasets

Our dataset can be represented as $X = \{x_1, x_2, ..., x_N\}$ where N is the total number of hospital visits. It is composed of multi-variate irregular time series data in that a visit x_i consists of a sequence of events: $x_i = \{x_i^1, ..., x_i^{T_i}\}$, where x_i^t represents the patient's records at time step t in x_i. We have $x_i^t \in \mathbb{R}^D$, where D is the number of attributes/features recorded at each event and T_i is number of events in the visit i which varies with different visits.

For each x_i, we have two levels of labels: the visit-level label for entire x_i and the event-level label for each of $\{x_i^1, ..., x_i^{T_i}\}$ respectively. At visit level, we use ICD-9 and thus we have $Y = \{y_1, y_2, ..., y_N\}$. At event level, we have corresponding labels $y_i = \{y_i^1, ..., y_i^{T_i}\}$ for the sequence of events $x_i = \{x_i^1, ..., x_i^{T_i}\}$ by using the following clinical experts designed criteria:
$y_i^t = 1$ (a patient is in septic shock state at a given time t in the visit x_i) if:

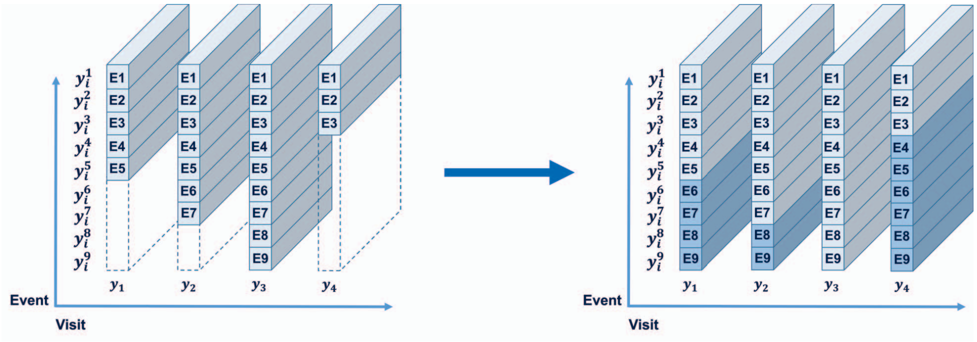


Figure 1: A toy example of data and padding procedure. The dark parts are the padded events to make each visit at the same length.

- Persistent hypotension as shown through two consecutive readings of:
  - SBP (systolic blood pressure) $<$ 90 mmHg;
  - or MAP (mean arterial pressure) $<$ 65 mmHg;
  - or SBP $\geq$ 40 mmHg below baseline.
- Or vasopressor administration.

Otherwise, $y_i^t = 0$.

For simplicity, we omit index i hereinafter when it is clear.

A. *Data Description*

Our dataset consists of anonymized clinical multi-variate time series data extracted from EHR system at Christiana Care Health System. Each data point is a visit/episode constituted by several events. In total there are 12,980 visits and 1,446,225 events. Besides identifiers and description information, each event consists of critical attributes/features into 4 categories:

- Vital signs: systolic blood pressure, diastolic blood pressure, mean arterial pressure, temperature, heart rate, respiratory rate, pulse oxygen.
- Lab results: BUN, procalcitonin, platelet, creatinine, bilirubin, C-reactive protein, white blood cell count, Bands, lactate, glasgow coma scores, glasgow best verbal response score, sedimentation rate, antibiotic/antiviral/anti-fungal drug administration, vasopressor administration.
- Location: location type, code and description. Specifically, there are 4 types of locations, emergency department, nurse, step down and intensive care unit (ICU) respectively.
- Oxygen: FIO2, oxygen flow, oxygen source.

Additionally, each event includes 18 culture test results.

B. *Data Preprocessing*

Each visit consists of irregular multi-variate time series events with missing values and missing attributes. This is caused by different attributes are measured at different events. For example, vital signs are measured every 8 hours while lab values are measured only every 24 hours. Hence there doesn't always exist available readings for lab results when a new event is created for vital signs. To fill the gaps, the values of vital signs are carried forward for 8 hours and the values of lab results are carried forward for 24 hours, guided by clinicians.

The number of events in a visit ranges from three to 3,464. Among them, 99.8% visits have less than 584 events. For the ease of computing in LSTM models, 26 visits with more than 584 events, are excluded in our following experiments. The remaining 12,954 visits spans a time period from three minutes to 65 days. Additionally, we padded short sequences with zeros so that each visit is a sequence of 584 events. Consider a toy example of data and padding procedure in Figure I: originally four slices of data/visits have five, seven, nine and three events respectively and after padding, all visit consists of nine events.

C. *Data Separation and Ground-truth*

Table I: The distribution of visits (events) with respect to the agreement between visit-level label and event-level label.

Label Level	**ICD-9** $(+)$	**ICD-9** $(-)$
Event $(+)$	$1,611(277,421)$	$1,492(248,737)$
Event $(-)$	$327(48,813)$	$9,524(773,654)$

Table I shows the distribution of visits (events) regarding the agreement between the two levels of labels in our dataset. The two-level labels are agreed (in shade) if: 1) for positive samples, ICD-9 indicates a septic shock visit and there exists at least one event in this visit meeting the shock criteria (i.e. with positive event-level label); and 2) for negative samples, ICD-9 shows no indication of septic shock and all events in this visit are not labeled as shock by the event-level criteria. In the following, the visits and events belong to these two shaded cells are referred as agreed data. As shown in Table I, the original data contain both agreed and disagreed samples. While they agree on the majority of the data, there are substantial amount of the data, about 14% of visits, the two levels of the labels disagree. Especially, note that for all the visits with at least one positive shock label at event level, more than 48% are labeled as non-shock by ICD-9 code.

To evaluate our proposed framework, we need to validate it on dataset with reliable labels. To do so, 80% of the *agreed data* are selected as our **test set** and our **training data** are selected from the remaining 20% agreed data and all the disagreed data through stratified random sampling.

Since the labels of the test set are used for evaluating our framework, their labels need to be highly reliable and reasonable as ground-truth. Thus they are always adjusted according to one of the following three ways:

- ***Prior*$_6$*:*** re-label all the events in the prior 6 hours of any shock event as positive; or
- ***Post*$_8$*:*** re-label all the events in the following 8 hours of any shock event as positive; or
- ***Between*$_8$*:*** re-label all events between any two shock events as positive if their interval is less or equal to 8 hours.

For training dataset, we have the options of either adjusting the labels in the same procedure as the test data or keeping it as it is. In the following we run separate experiments on learning a model using training set with original labels or with the adjusted labels. We refer to the latter as the adjusted training set as 'train$_{adj}$'. By testing on both original training set and 'train$_{adj}$', we can inspect how much LSTM could learn without manual adjustment and whether the proposed framework can contribute more than manual adjustment.

III. Method

In this section, we start with the description of the traditional Long-Short Term Memory networks (LSTMs). Then we demonstrate how to combine two levels of imperfect labels, i.e. the visit-level labels and the event-level labels in a unified framework to predict septic shock.

A. LSTM for sequential classification

Sequential data in real applications are often collected over a long span of time, such as health-care records, climate changes, traffic monitoring, etc. In our problem, the longest visit is over two months. Therefore we utilize LSTM model in detecting septic shock for its capacity of capturing long-term sequential patterns.

We first briefly introduce the LSTM model, which is a variant of Recurrent Neural Networks (RNN). Rather than conducting classification at each time step separately, LSTM introduces a LSTM cell to model the transitions between different time steps. The structure of LSTM cell is shown in Fig. 2(a). Each LSTM cell contains a cell state c^t, which serves as a memory and controls the information flow, added, removed or unchanged. Each LSTM cell at t also outputs a hidden representation h^t, which is a high-level representation of information at current time step, and can be used for classification at t.

The cell state c^t is generated by combining $\{c^{t-1}$,$h^{t-1}\}$ and the information at t. Specifically, we first decide what information should be added to the current cell state by generating a new candidate cell state. We also generate an input gate to filter the new added information. The gating variables in LSTM cell, e.g input gate, are computed by a sigmoid function with combination of x^t and h^{t-1}, while the candidate cell state is generated by a $tanh(\cdot)$ function, as follows:

$$\begin{aligned} i^t &= \sigma(W_h^i h^{t-1} + W_x^i x^t), \\ \tilde{c}^t &= tanh(W_h^c h^{t-1} + W_x^c x^t), \end{aligned} \tag{1}$$

where $\{W_h^i \in \mathbb{R}^{H\times H}, W_x^i \in \mathbb{R}^{H\times D}\}$ and $\{W_h^c,\ W_x^c\}$ denote two sets of weight parameters for generating the input gate and candidate cell state respectively. Hereinafter we omit the bias terms as they can be absorbed into weight matrices. Next, we create a forget gate f^t using a sigmoid function to remove information from the past:

$$f^t = \sigma(W_h^f h^{t-1} + W_x^f x^t), \tag{2}$$

where $\{W_h^f,\ W_x^f\}$ denotes the weight parameters used to generate the forget gate layer f^t.

In the way of forgetting old information from the old state c^{t-1} and filtering new information from the candidate cell state at time t, we obtain the new cell state c^t as follows:

$$c^t = f^t \otimes c^{t-1} + i^t \otimes \tilde{c}^t, \tag{3}$$

where $\otimes$ denotes entry-wise product.

Finally, we generate the hidden representation by filtering the the obtained new cell state by an output gate layer o^t:

$$\begin{aligned} o^t &= \sigma(W_h^o h^{t-1} + W_x^o x^t), \\ h^t &= o^t \otimes tanh(c^t), \end{aligned} \tag{4}$$

where $\{W_h^o, W_x^o\}$ are weight parameters that are used to generate the output gate layer. The output gate determines the information to output from c^t to h^t.

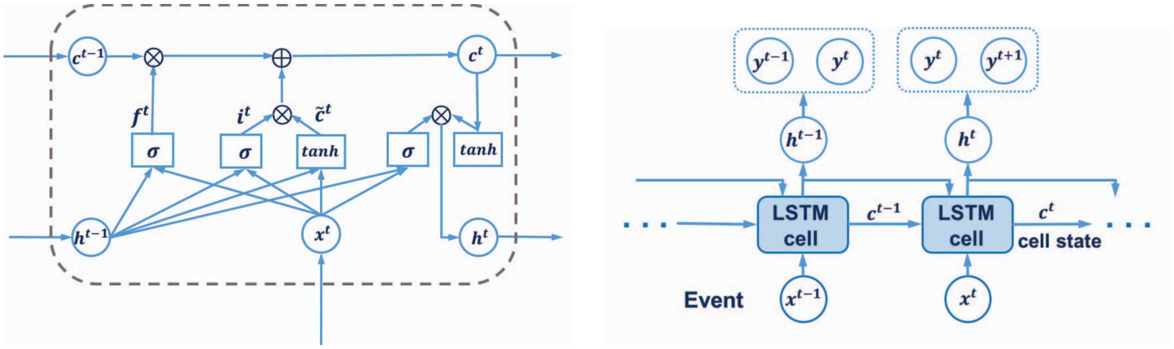
With the hidden representation h^t, we produce the probability of each event t at risk of septic shock using a sigmoid function with parameter U, as follows:

$$p^t = \sigma(Uh^t). \tag{5}$$

B. Learning from multi-level labels

We have two levels of labels for septic shock, ICD-9 code as visit-level label and event-level labels generated by clinical criteria. Even though informative for septic shock progression, both of them are imperfect and contain false positives and false negatives. Therefore we propose to combine two sources of labels to the learning process of LSTM to jointly learn the sequential patterns of septic shock.

Basically, we reinforce the learning process by revising the provided labels with confidence. The LSTM model is



(a)

(b)

Figure 2: (a) The structure of LSTM cell. (b) The structure of the proposed sequential classification method for early prediction, a variant of LSTM.

firstly trained on the original event-level data. After training, the model grasps the event-level septic shock patterns and generates the probabilities of each event at the shock state, p_i^t. Then the predicted output labels, $\hat{y}_i^t$, are sampled according to p_i^t. Note that this output is different from original event-level labels in that it is generated by the model according to its learned shock patterns from training data.

With the obtained event-level output probabilities, we then check the consistency between event-level and visit-level information. Specifically, if the visit-level label y_i and the event-level output probabilities $\{p_i^t\}$ consistently indicate the same prediction results, i.e. shock or not, on some visits, event-level labels on those visits would be set accordingly to reinforce the learning objectives for the model. More formally, in a visit which is labelled as positive by ICD-9, if there exist m consecutive events at time steps $\{t_r, ..., t_r + m - 1\}$, whose sum of the output probabilities is bigger than a predetermined threshold $thr_1 \times m$, this visit is identified as a consistent positive visit. Then we find the consecutive m events with largest summed probabilities, denoted by $\{t_r^*, ..., t_r^* + m - 1\}$ and set the labels $\{y_i^{t_r^*}, ..., y_i^{t_r^*+m-1}\}$ as $\{1, ..., 1\}$. On the other hand, in a visit which is labelled as negative by ICD-9, if any m consecutive events, whose sum of the output probabilities is less than another predetermined threshold $thr_2 \times m$, this visit is identified as a consistent negative visit. . Then the labels of all the events in this visit are set to zero, i.e. $\{y_i^t\} = \{0\}$.

Then the LSTM model is updated using the data with revised labels by back-propagation algorithm. We repeat this process until no change is made to the labels and the model is convergent. The process of revising labels is only conducted every $epochs\_update$ iterations. The complete learning process is summarized in Algorithm 1.

Algorithm 1: Learning process for proposed framework with LSTM.

```
Input:
$\{x_i^1, x_i^2, ..., x_i^T\}_{i=1}^N$: A set of multi-variate sequences.
$\{y_i^1, y_i^2, ..., y_i^T\}_{i=1}^N$: Event-level labels of sequences.
$\{y_1, ..., y_N\}$: Visit-level labels of sequences.
Initialize parameters;
while not converge & $it$++ $\leq MaxIter$ do
    for $i \leftarrow 1$ to $N$ do
        // In practice we use mini-batch update.
        for $t \leftarrow 1$ to $T$ do
            Compute latent output $p_i^t$ by Eqs. 4 and 5;
            Compute the predicted label $\hat{y}_i^t$;
        Update parameters for one iteration by
          back-propagation;
    if mod(it,epochs_update)==0 then
        for $i \leftarrow 1$ to $N$ do
            if $y_i = 1$ & $\exists t_r, \ni \sum_{t=t_r}^{t_r+m-1} p_i^t > thr_1 \times m$
              then
                $t_r^* = \mathrm{argmax}_t \sum p_i^{t:t+m-1}$
                if $\exists t \in \{t_r^*, ...t_r^* + m - 1\} \ni \hat{y}_i^t! = y_i^t$ then
                    $y_i^t = 1$
            if $y_i = 0$ & $\forall t_r, \ni \sum_{t=t_r}^{t_r+m-1} p_i^t < thr_2 \times m$
              then
                if $\exists t \in \{1, ..., T\} \ni \hat{y}_i^t! = y_i^t$ then
                    $y_i^t = 0$
```

In our implementation, we set the value of m in Algorithm 1 as two, which stands for the two consecutive events. This value is concluded from the data and confirmed by domain experts. Figure 3 shows the distribution of the longest sequence of consecutive ones in non-shock and shock visits from perspective of ICD-9 code respectively, where the horizontal axis is the percentile and the vertical axis indicates the length of the longest sequence of consecutive ones in a visit. It is observed that in the non-shock visits identified by ICD-9 code, there are less than ten percent of visits have one shock event. Conversely, in the shock visits, roughly 80 percent of visits have at least one shock event and 70 percent of visits have at least two consecutive shock events. In practice, clinicians usually regard a visit as positive as long as there exists one positive event. Since the labels need to be carefully revised, it is more conservative to associate a positive visit with two consecutive positive events and no consecutive positive events would more safely indicate a negative visit.

Moreover, the framework can also be applied for early prediction that detects the patients 'trending septic shock'. To be specific, we utilize the future diagnostic condition as part of training labels for each current event in training process. For example, we combine future event's label y^{t+1} together with the current event's label y^t as the training objectives for the current event, as shown in Fig. 2(b). We still include y^t here since we usually also care about the current condition even in the early prediction task. In this way the model is expected to capture the important transitions between medical states of patients and be more sensitive about the progression of the disease.

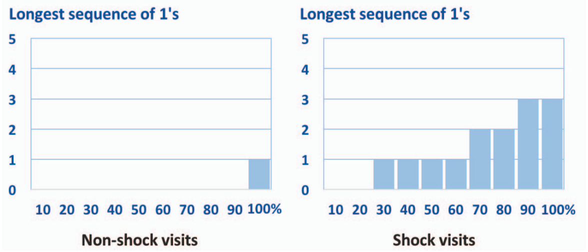


Figure 3: The distribution of consecutive 1's in non-shock and shock visits.

IV. Experiments

This section describes our evaluations and presents our results. First, we compare the performance of the traditional LSTM to the LSTM integrated with our framework that can leverage multi-level labels. Second, to explore the generality of our proposed framework, we combine it with other popular machine learning approaches for the task of septic shock prediction and show that the framework reaches the best performance when combining with LSTM. Third, to demonstrate the robustness of the proposed framework, we compare the prediction results of the framework with LSTM using three different sets of ground-truth labels - $Prior_6$, $Post_8$, and $Between_8$. Finally, we conduct experiments to test the effectiveness of the proposed method in early septic shock prediction.

The evaluation metrics include sensitivity/recall, specificity/true negative rate (TPR), precision/positive predictive value (PPV), negative predictive value (NPV), F1 score, and area under the ROC (receiver operator characteristic) curve (AUC) [21]. Precision, recall, F1-score and AUC are widely used to measure the prediction performance for machine learning approaches. In medical science domain, researchers commonly refer to sensitivity (i.e. recall), specificity (i.e. TPR) and PPV for the annotation performance. Therefore in our tests, we include the metrics for both machine learning and medical science domains. In the learning process, we randomly take 30% of training data as validation set for hyper-parameter tuning. As described in Section II-C, the labels of test data are always adjusted according to one of three ways used as ground-truth, and training set is in both original form and adjusted with the same procedure as test data denoted as 'train$_{adj}$'. For all the tests in this section, we repeat each experiment 10 times with random initialization and random selection of validation set and report the average values and standard deviation for each evaluation metric.

A. Validation of Framework

The effectiveness of the proposed framework is evaluated by comparing the performances of LSTM and LSTM^f (LSTM with the proposed framework) on both train and train$_{adj}$. Aforementioned we have three sets of constructed ground-truth labels. Given the space limitation, here we only report the experimental results on $\boldsymbol{Between}_8$, which is widely accepted by clinicians and labelling all the events between two shock events within the period of 8 hours as positive events. The similar results are reflected when using the other two types of clinical adjustments.

After running the grid search we set the number of hidden units for LSTM is 24. Each model is trained on 200 epochs and the labels are updated every 10 epochs ($epochs\_update$) in our proposed framework. Furthermore, the thresholds in Algorithm 1 are $thr_1 = 0.7$ and $thr_2 = 0.5$ respectively.

Table II shows the results of comparing four types of training models on our test dataset. The four models are learned by applying either LSTM or LSTM^f on the original train dataset or train$_{adj}$ respectively. Overall, the combination of train$_{adj}$ data and LSTM with framework ($\{$train$_{adj}$, LSTM$^f\}$) achieves the best performance, which has the highest values of recall, NPV, AUC and F1 score. The other two evaluation metrics of this model, specificity and precision, are only about 2% lower than the highest ones reached by the model $\{$train, LSTM$\}$. It is noticeable that the variation range of specificity and precision is as minor as two percent. Conversely, the other four metrics, recall, NPV, AUC and F1 are greatly improved. For example, sensitivity increases 61% from 0.619 ($\{$train, LSTM$\}$) to 0.996 ($\{$train$_{adj}$, LSTM$^f\}$) and even NPV increases 10% from 0.907 approaching to a perfect one. Also, the rank of the models under these four evaluation metrics are consistent, where $\{$train$_{adj}$, LSTM$^f\}$ is the best, followed by $\{$train, LSTM$^f\}$, $\{$train$_{adj}$, LSTM$\}$ and $\{$train, LSTM$\}$.

Table II shows that $\{$train$_{adj}$, LSTM$\}$ greatly outperforms $\{$train, LSTM$\}$ on several important metrics including recall, AUC and F1 score. Such result is not surprising in that the LSTM learns better from the training data that is more similar to test data since both train$_{adj}$ and test data are processed with the same clinical adjustment. On the other hand, this result also shows that LSTM indeed captures the temporal dependencies that exist in the adjusted labels.

Table II: Performance(±standard deviation) LSTM with or without the framework.

	Sensitivity/Recall	**Specificity**	**PPV/Precision**	**NPV**	**AUC**	**F1**
{train, LSTM}	0.619(±0.018)	**0.975**(±0.011)	**0.867**(±0.018)	0.907(±0.012)	0.723(±0.014)	0.722(±0.015)
{train$_{adj}$, LSTM}	0.842(±0.011)	0.962(±0.008)	0.854(±0.013)	0.959(±0.008)	0.902(±0.009)	0.848(±0.010)
{train, LSTM^f}	0.947(±0.015)	0.958(±0.011)	0.855(±0.016)	0.986(±0.011)	0.985(±0.019)	0.897(±0.014)
{train$_{adj}$, LSTM^f}	**0.996**(±0.003)	0.954(±0.019)	0.851(±0.020)	**0.999**(±0.001)	**0.988**(±0.008)	**0.918**(±0.012)

⋆ The best number within each metric across the four models is in bold.

Table III: Performance(±standard deviation) of multiple baselines with framework.

	Sensitivity/Recall	**Specificity**	**PPV/Precision**	**NPV**	**AUC**	**F1**
{train$_{adj}$, LR^f}	0.359(±0.028)	**0.999**(±0.001)	**0.994**(±0.004)	0.856(±0.012)	0.679 (±0.026)	0.527(±0.018)
{train$_{adj}$, SVM^f}	0.436(±0.011)	0.987(±0.010)	0.899(±0.021)	0.870(±0.008)	0.712(±0.015)	0.587(±0.010)
{train$_{adj}$, RNN^f}	0.608(±0.021)	0.976(±0.013)	0.868(±0.014)	0.905(±0.011)	0.951(±0.017)	0.715(±0.016)
{train$_{adj}$, GRU^f}	0.938(±0.015)	0.958(±0.009)	0.856(±0.012)	0.983(±0.009)	0.982(±0.011)	0.895(±0.014)
{train$_{adj}$, LSTM^f}	**0.996**(±0.003)	0.954(±0.019)	0.851(±0.020)	**0.999**(±0.001)	**0.988**(±0.008)	**0.918**(±0.012)

⋆ The best number within each metric across the five approaches is in bold.

Similarly, the comparison between {train, LSTM^f} and {train$_{adj}$, LSTM^f} shows the latter outperforms the former. Moreover, compared with the improvement from {train, LSTM} to {train$_{adj}$, LSTM}, relatively smaller improvement was found from {train, LSTM^f} to {train$_{adj}$, LSTM^f}. This indicates that our proposed framework is less sensitive to the quality of the training data probably because the proposed framework can truly learn from both levels of informative yet imperfect labels.

Additionally, the comparison of {train, LSTM} against {train, LSTM^f}, and {train$_{adj}$, LSTM} against {train$_{adj}$, LSTM^f} further demonstrate the effectiveness of our proposed framework. When learning with the original training data, recall increases 53% from 0.619 to 0.947, AUC increases 36% from 0.723 to 0.985, and F1 increases 24% from 0.722 to 0.897. When learning with the adjusted training data, the improvements are still considerable. For instance, the improvements for recall, AUC and F1 are 18%, 10% and 8%, respectively.

As a brief summary, in the training process, our proposed framework can effectively update the learning objectives to approach the latent true labels of the data and then achieve a better performance. By jointly extracting useful information from two levels of labels, the model learns the discriminative patterns of septic shock and utilizes these patterns to automatically identify noisy labels.

B. Generality of the Proposed Framework

To explore the generality of our proposed framework, we incorporate it with other popular machine learning approaches for the task of septic shock prediction. Such popular approaches include two classic approaches: Logistic Regression (LR) and Support Vector Machine (SVM), which are non-sequential models and are ubiquitously implemented due to their efficiency and robustness, and two temporal sequential models: Recurrent Neural Networks (RNN) and Gated Recurrent Units (GRUs). Both LSTM and GRU are special variants of RNN. GRU, which has gained popularity recently, can also be treated as a simpler version of LSTM.

For SVM baseline, we adopt RBF kernel, which is justified by the optimal performance in validation set. For a fair comparison, we set RNN and GRU to share the same model settings with LSTM, e.g. the number of epochs and hidden units. The performance of these methods using the proposed framework is shown in Table III.

Table III shows that LR^f obtains the best specificity and precision while LSTM^f obtains the best performance on all the remaining four evaluation metrics. Generally, the ordering from low performance to high performance among these approaches is the same as the listed order in Table III. In particular, while NPV and AUC of LR and SVM are acceptable, their recalls are very low, under half of recall values achieved by LSTM^f. As a result, their F1 scores are too low to be accepted for the detection task (as it is not allowed to miss too many true alarms of septic shock). This is largely because LR and SVM cannot capture the septic shock progression because both models do not take account of the temporal dependencies among the sequential events. On the other hand, the improvements made by GRU and LSTM over RNN demonstrate that incorporating the long-memory structure in the model greatly helps capture the underlying patterns of septic shock. Finally, the results of GRU and LSTM are very close with each other.

To summarize, sequential models are more appropriate for septic shock prediction since they are considering the sequential dependencies among the events, which is extremely useful for reasoning the development of septic shock. Compared with traditional RNN, GRU and LSTM are more preferable since they are capable of handling long-term dependencies and persistently keep the important long-term underlying patterns for future reference, e.g., the septic shock usually persists for a period.

Table IV: Performance($\pm$standard deviation) of the framework on test data with different ground-truth labels.

Method	Sensitivity/Recall	Specificity	PPV/Precision	NPV	AUC	F1
{train$_{adj}$, LSTM^f, *Prior*$_6$}	0.999($\pm$0.001)	0.967($\pm$0.012)	0.926($\pm$0.021)	0.999($\pm$0.001)	0.990($\pm$0.004)	0.961($\pm$0.013)
{train$_{adj}$, LSTM^f, *Post*$_8$}	0.999($\pm$0.001)	0.958($\pm$0.014)	0.882($\pm$0.022)	1.000($\pm$0.000)	0.989($\pm$0.009)	0.937($\pm$0.011)
{train$_{adj}$, LSTM^f, *Between*$_8$}	0.996($\pm$0.003)	0.954($\pm$0.019)	0.851($\pm$0.020)	0.999($\pm$0.001)	0.988($\pm$0.008)	0.918($\pm$0.012)

C. Validation Using Different Ground-truth Labels

In practice it is very hard to collect and confirm the ground truth of septic shock. There exist multiple reasonable ways to adjust labels to approach the ground-truth, which usually generate different ground-truth sets. Here we implement our proposed framework on three different sets of ground-truth labels to testify its robustness. The three ground-truth labels of the test data are produced by following the consultation of clinicians to adjust the labels as described in section II-C, which are *Prior*$_6$, *Post*$_8$ and *Between*$_8$.

Since the settings {train$_{adj}$, LSTM^f} lead to the best performance according to Table II, in this section we focus on using LSTM^f and measure its performance with three different types of adjusted labels for both training labels and test labels. The results of the proposed framework on three groups of ground-truth labels are listed in Table IV. It can be observed that our proposed LSTM^f works well on all three types of clinical adjustment. More specifically, the values of recall, specificity, NPV and AUC of the *Prior*$_6$ and *Post*$_8$ are as good as those of *Between*$_8$. On the other hand, these two groups achieves even higher precision and F1 score than *Between*$_8$. Among three groups, the adjustment on *Between*$_8$ is most conservative since it labels positive events only when two positive events happen within eight hours. The results confirm that on all three groups of ground-truth labels, our proposed framework can timely and effectively learn the progression of septic shock and also discover latent true labels even from imperfect labels.

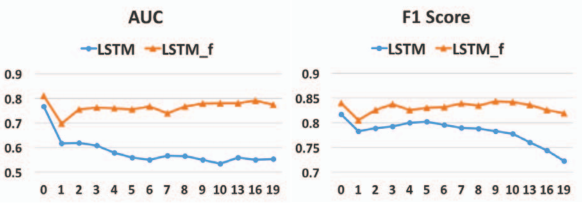
D. Early Prediction

In this section, we investigate the effectiveness of the proposed framework for early prediction, i.e. to detect the patients who are developing septic shock in advance.

In our implementation, we utilize the future diagnostic condition as part of training labels for each current event in training process, as described in Section III-B. Specifically, we shift labels of prospective events to present events and combine with the label of the current event. For example, if we wish to predict η events in advance, the original label of a event y^t is changed to $\{y^t, y^{t+\eta}\}$. Training data and testing data are processed with the same shifting procedure and then used for training and evaluating LSTM^f.

For early prediction, it is extremely important to capture the positive class, such as the label pair of $\{0, 1\}$ and $\{1, 1\}$. Our data is heavily skewed with large volume of $\{0, 0\}$, which accounts for more than 60% of the data. To fairly reflect the performance of our framework on septic shock detection, we keep all positive events and randomly sample similar amount of negative events. Here we show the performance using F1 and AUC, as F1 score computes a balanced measure for positive class, and AUC reflects the relationship between false positives and true positives.

In this part we conduct experiments on the model of adjusted training data and ground-truth data (***Between***$_8$) for cases either with or without the proposed framework. We track the performance of each method from one-event early up to 19-event early. In particular, the performance is recorded every time before 10-event early and after that the performance is recorded every three-event early. The variation of the two metrics with shifting labels are shown in Figure 4. Note that the starting point '0' stands for general prediction of current event without any shifting processes.



(a) (b)

Figure 4: (a) Variation of AUC of the framework with shifting labels (b) Variation of F1 score of the framework with shifting labels.

Overall, LSTM^f is always better than the traditional LSTM, and both models have their performance gradually degraded when predicting future event labels. However, the integration of the framework can mitigate the declining trends. It is obvious LSTM^f is more stable and decreases much slower in terms of both AUC and F1 metrics. While the mapping relation from features to prospective labels is difficult to learn, the temporal progression patterns can still be helpful for a reasonable prediction. By inferring true labels using our proposed framework, the LSTM model is more likely to capture the temporal septic shock progression using the revised event-level labels. When implemented without using the proposed framework, AUC of LSTM drops from 0.767 to 0.534 and F1 score is from 0.817 to 0.720. It is worth noticing that the AUC of LSTM quickly shrinks to around 0.550, as well as a clear and faster declining trend in F1. The results indicate that the proposed framework can effectively and stably detect septic shock at an early stage with high performance.

V. Related Works

This literature review provides the related works to the proposed interdisciplinary research between machine learning and health informatics. Here we will also cover multiple sepsis stages, e.g. sepsis, severe sepsis and septic shock.

There exist plenty of general-purpose illness severity scoring systems to access illness severity and risk of death among septic patients, such as Modified Early Warning Score (MEWS), Simple Clinical Score (SCS), mortality in emergency department sepsis (MEDS), rapid emergency medicine score (REMS), the Acute Physiology and Chronic Health Evaluation (APACHE II), Simplified Acute Physiology Score (SAPS II) [9] [1] . Even though these scores have been validated to be useful for general deterioration and mortality, they are not specific for sepsis and cannot distinguish patients at highest risk of developing an acute condition with high sensitivity and specificity [13]. In specific, the AUC of SCS and REMS, which are concluded as the most appropriate clinical scores to predict the mortality of patients with sepsis, are 0.76-0.79 and 0.74-0.79 respectively.

Prior studies have shown that early diagnosis and treatment of septic shock can decrease patients mortality and length of stay significantly [18] [5] [13]. This stimulates the development of automating tools, e.g. early warning systems, track and trigger initiatives, and listening applications [14] [22] [31] , which enables a timely therapy for patients. Although these tools can successfully detect patients who are experiencing septic shock, they are not able to predict if patients are at greatest risk of developing septic shock.

Most recent studies have concentrated on early prediction of septic shock with the increasing popularity of EHR data [10] [29] [12] . Machine learning techniques are widely applied to construct predictive models by routine measurements [11] . For e.g., a septic shock early warning system (EWS) is developed using multivariate logistic regression to predict septic shock one hour prior to onset, and they achieved an AUC of 0.940, a sensitivity of 0.85 and a PPV of 0.70 [25]. A targeted real-time early warning score (TREWScore) is fit in a Cox proportional hazards model and reports an AUC of 0.83, specificity of 0.85 and sensitivity of 0.74 [13]. Peelen *et al.* [24] developed a dynamic Bayesian network to model the progression of organ failure, which reaches an AUC of 0.911 24-hour after admission and 0.944 using the training data of first three-hour. Additionally, SVM has proven to be effective at sepsis-related classification tasks [32] [28]. but there exists no quantification in terms of time on how early of the prediction.

LSTM has shown extensive prospects in a variety of sequential labeling applications, such as climate changes, health-care records, traffic monitoring, etc [19] [17] [16] [8]. Originally, neural networks have been applied to medical problems for more than 20 years [2] [3] [34]. In particular LSTM has been implemented for general diagnose termed as phenotyping [19] [8]. RNN, the archetype LSTM inherits from, has been applied to the field of medical care such as physiologic signals and prediction problems in genomics [33].To our best knowledge there is no work on specifically applying LSTM to the task of sepsis prediction.

For the learning process, prior works only employ one-level label. Some studies take the intersection of ICD-9 code and event-level criteria to select positive and negative cases, e.g., septic shock patients are identified by ICD-9 code and the need for vasopressors within 24 hours of ICU transfer [29]. A popular but non-specific indicator, Systemic Inflammatory Response Syndrome (SIRS), together with ICD-9 are used to screen sepsis cases [4]. However, none of them combine two levels of labels to jointly learn the patterns of septic shock.

VI. Conclusion

In this work, we propose a framework to combine two unreliable sources of labels, visit-level label and event-level label to reliably predict septic shock on a real-world EHR dataset. The framework is implemented on a variant of LSTM model to capture the temporal progression of septic shock during a long visit. The experimental results demonstrate the superiority of the proposed framework and the effectiveness of LSTM compared with multiple baselines. Its robustness is also validated on the test data with three different ground-truth labels. Besides, the proposed method can not only detect patients in the shock state, but also stably predict patients who are trending septic shock.

Here we investigate our future works in two aspects. First, we will quantify the effectiveness of early prediction in terms of concrete time measure, thus to better assist clinicians in making decisions. Second, we will collaborate with Mayo Clinic and implement the proposed method on larger EHR datasets, which can better validate the general performance of the proposed method.

Acknowledgment

This work was funded by the NSF Awards #1522107 and #1651909.

Reference

[1] Steven L Barriere and Stephen F Lowry. An overview of mortality risk prediction in sepsis. *CCM*, 1995.

[2] William G Baxt. Application of artificial neural networks to clinical medicine. *The lancet*, 1995.

[3] R Brause, F Hamker, and Jürgen Paetz. Septic shock diagnosis by neural networks and rule based systems. *Studies in Fuzziness and Soft Computing*, 2002.

[4] Jacob S Calvert, Daniel A Price, Uli K Chettipally, et al. A computational approach to early sepsis detection. *Computers in biology and medicine*, 2016.

[5] Victor Coba, Melissa Whitmill, et al. Resuscitation bundle compliance in severe sepsis and septic shock: improves survival, is better late than never. *Journal of intensive care medicine*, 2011.

[6] 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.

[7] Anne Elixhauser, Bernard Friedman, and Elizabeth Stranges. Septicemia in U.S. hospitals, 2009. https://www.ncbi.nlm.nih.gov/books/NBK65391/, 2011.

[8] Cristóbal Esteban, Antonio Artés, Yinchong Yang, Oliver Staeck, Enrique Baca-Garcıa, and Volker Tresp. Combining static and dynamic information for clinical event prediction.

[9] Nesrin O Ghanem-Zoubi, Moshe Vardi, Arie Laor, Gabriel Weber, and Haim Bitterman. Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments. *Critical Care*, 2011.

[10] Karen K Giuliano. Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis. *American Journal of Critical Care*, 2007.

[11] Eren Gultepe, Jeffrey Green, et al. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *JAMIA*, 2014.

[12] Katharine Henry, Chris Paxton, Kwang Sik Kim, Julius Pham, and Suchi Saria. 63: Rews: Real-time early warning score for septic shock. *Critical Care Medicine*, 2014.

[13] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *STM*, 2015.

[14] Vitaly Herasevich, Matthew S Pieper, Juan Pulido, and Ognjen Gajic. Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation. *JAMIA*, 2011.

[15] Joyce C Ho, Cheng H Lee, and Joydeep Ghosh. Septic shock prediction for patients with missing data. *ACM Transactions on Management Information Systems (TMIS)*, 2014.

[16] 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.

[17] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Predict land covers with transition modeling and incremental learning. In *Proceedings of the 2017 SDM*, 2017.

[18] Anand Kumar, Daniel Roberts, Kenneth E Wood, 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.

[19] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[20] Greg S Martin, David M Mannino, Stephanie Eaton, and Marc Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 2003.

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

[22] Su Q Nguyen, Edwin Mwakalindile, et al. Automated electronic medical record sepsis detection in the emergency department. *PeerJ*, 2014.

[23] U.S. Department of Health and Human Services. Health, United States, 2016: With chartbook on long-term trends in health. https://www.cdc.gov/nchs/data/hus/hus16.pdf#019, 2017.

[24] Linda Peelen, Nicolette F de Keizer, Evert de Jonge, et al. Using hierarchical dynamic bayesian networks to investigate dynamics of organ failure in patients in the intensive care unit. *Journal of biomedical informatics*, 2010.

[25] Dewang Shavdia. *Septic shock: Providing early warnings through multivariate logistic regression models*. PhD thesis, Massachusetts Institute of Technology, 2007.

[26] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). 2016.

[27] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[28] Collin HH Tang, Paul M Middleton, et al. Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. *Physiological measurement*, 2010.

[29] Steven W Thiel, Jamie M Rosini, William Shannon, Joshua A Doherty, Scott T Micek, and Marin H Kollef. Early prediction of septic shock in hospitalized patients. *JHM*, 2010.

[30] CM Torio and RM Andrews. National inpatient hospital costs: the most expensive conditions by payer, 2013. rockville, md: Agency for healthcare research and quality; 2013, 2016.

[31] Craig A Umscheid, Joel Betesh, Christine VanZandbergen, et al. Development, implementation, and impact of an automated early warning and response system for sepsis. *Journal of hospital medicine*, 2015.

[32] Shu-Li Wang, Fan Wu, and Bo-Hang Wang. Prediction of severe sepsis using svm model. In *Advances in computational biology*. Springer, 2010.

[33] Rui Xu, Donald Wunsch II, and Ronald Frank. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE TCBB*, 2007.

[34] Guangxu Xun, Xiaowei Jia, and Aidong Zhang. Detecting epileptic seizures with electroencephalogram via a context-learning model. *BMC medical informatics and decision making*, 2016.