Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information using Convolutional-LSTM

Chen Lin*, Yuan Zhang†, Julie Ivy ‡, Muge Capan §,Ryan Arnold ¶, Jeanne M. Huddleston ‖, and Min Chi**
*†‡** Department of Computer Science, North Carolina State University, Raleigh, NC, USA
§ Decision Sciences & MIS Department, Drexel University, Philadelphia, PA, USA
¶ Department of Emergency Medicine, Drexel University, Philadelphia, PA, USA
‖ Mayo Clinic, Rochester, MN, USA
Email: *clin12@ncsu.edu, †yzhang93@ncsu.edu, ‡jsivy@ncsu.edu, §Muge.Capan@drexel.edu,
¶ryanarnold08@mac.com, ‖huddleston.jeanne@mayo.edu, and **mchi@ncsu.edu

***Abstract*—Deep neural network models, especially Long Short Term Memory (LSTM), have shown great success in analyzing Electronic Health Records (EHRs) due to their ability to capture temporal dependencies in time series data. In this paper, we proposed a general deep neural network framework which incorporates two additional components with the aim of improving LSTM. The first component, a Convolutional Neural Network (CNN), is added before LSTM to obtain local characteristics of EHRs. The second component, a fully connected neural network (FC), introduces static information (e.g., age) to LSTM, which is applied to handle dynamic information (e.g., lab result). The medical condition we aim to predict is septic shock – it is the most advanced complication of sepsis and is due to severe abnormalities in circulation and/or cellular metabolism. Our proposed framework was evaluated for two experimental tasks: *visit level early diagnosis* (left align) and *event level early prediction* (right align). Our results show that for *visit level early diagnosis*, by incorporating both CNN and static information, our framework consistently outperforms the original LSTM. For *event level early prediction*, the same outcome is observed when predicting $<$ 5 hours into the future, however, when predicting $\geq$ 5 hours into the future, the addition of the CNN component alone obtains the best results.**

***Index Terms*—Convolutional-LSTM, septic shock, Electronic Health Records, early prediction, early diagnosis**

I. Introduction

In recent years, an increasing amount of electronic health records (EHRs) [1] have become available. EHRs are systematic collections of patients' digital health records which include both static and dynamic information. The static information, such as gender, is collected only once per visit and remains unchanged for the duration of the visit; whereas dynamic information, such as lab tests, is collected multiple times during a patient's visit. Generally speaking, static information is represented vectorially while dynamic information is represented as a time series.

The adoption of EHRs has brought tremendous opportunities in developing data-driven machine learning models to support clinical decisions. For example, Markov models [2, 3, 4] and dynamic Bayesian Network [5] are commonly used to capture disease progression by modeling the temporal aspects of EHRs. However, Markov models assume a patient's current health state only depends on his previous health state. This is a very strong assumption and may not hold in all cases. Certain diseases may depend on a subset of the medical history or even its entirety. Thus, we proposed a different approach without making strong assumptions. The proposed framework is based on Long Short Term Memory (LSTM). The model was further extended by introducing two additional components: Convolutional Neural Network (CNN) and Fully Connected Neural Network (FC). Details of these three components are described below.

LSTM [6] is effective in capturing underlying temporal structures in time series data. It builds up memory by feeding the previous hidden state as an additional input into the subsequent step. This makes the model particularly suitable at modeling dynamic information in EHRs, where there is a strong statistical dependency between medical events over long-time intervals. This dependency is a key feature in recognizing early signs of physiological deterioration.

Conversely, CNN [7] has been shown to achieve remarkable results in many image/video related tasks [7, 8, 9] as it can effectively capture time-invariant features within short temporal regions – a task that may be more challenging for LSTM. These features are time-invariant as they can occur at any point in time and can be extracted from a portion of the records rather than relying on its entirety. Therefore, CNN is introduced before LSTM to extract local features from raw EHRs into compact state representations so that LSTM can recognize temporal dependency more easily.

Previous research suggests that static information such as comorbidities [10] and gender [11] are important predictors for septic shock. Therefore, our framework employs a fully connected neural network (FC) to incorporate static information into LSTM which only handles dynamic information. A study [12] sharing the same objective (i.e., incorporate static information) used outpatient data to predict the outcome of a patient's next visit. Note that for their dataset, the static information changes from visit to visit. As a result, they

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00032

concatenate the static information to the hidden states of LSTM at *every time step*. On the other hand, our study used inpatient data where the static information was collected only once in a visit and remained unchanged for the duration of that visit. Therefore, we concatenate the static information with LSTM only at *the last time step*. In our experiments, both approaches were explored.

The proposed framework was applied to septic shock prediction. Sepsis is a critical condition that arises when the body starts to injure its own tissues and organs during an infection [13]. Septic shock is the most severe stage of sepsis and results in a dramatic increase in the risk of death. In fact, mortality rate increases by 7.6% with each hour treatment is delayed after the onset of hypotension [14]. Therefore, early recognition and treatment is key to patient survival. General-purpose severity scoring systems such as Acute Physiology and Chronic Health Evaluation (APACHE II) [15], Simplified Acute Physiology Score (SAPS II) [16], and Sequential Organ Failure Assessment Scores (SOFA) [17] are used to predict general deterioration or mortality [18], but they cannot identify patients at high risk for septic shock with high sensitivity and specificity. Some tools can successfully identify patients who are already in septic shock but cannot accurately predict who will have septic shock in the future [18]. On the other hand, models such as the TREWScore, which do allow real-time identification of high-risk patients, are limited because they are unable to handle temporal sequential datasets well and rely heavily on manual feature engineering. To this day, predicting the onset of septic shock remains a challenge. Many patients at high-risk of septic shock are under-diagnosed at earlier stages of sepsis when aggressive treatment could still reverse its course of progression [19]. The ultimate goal of the proposed framework is to develop a predictive tool which allows patients at high risk of septic shock to be identified as early as possible.

Three major contributions of our work are:

- This framework demonstrates a novel neural network architecture specifically designed to handle static and dynamic information in EHRs. As far as we know, it is the first study that combines CNN, LSTM, and an independent FC to predict septic shock.
- This study explores the individual and combined efficacy of these two components (i.e., CNN and static information) by comparing the following architectures: original LSTM, LSTM combined with CNN only, LSTM combined with static information only and LSTM combined with both CNN and static information.
- This study proposed a novel experimental design exploring two disease prediction tasks: visit-level early diagnosis (left align) and event-level early prediction (right align). This experimental design provides a comprehensive examination of the model under distinct scenarios.

II. Related Work

Classic Machine Learning Models. Numerous classic machine learning models have been applied to tasks related to sepsis/septic shock, e.g., logistic regression [20]. Classic machine learning models do not adequately capture the temporal sequential nature of the medical events as well as their dependencies. These models were trained using summary statistics (e.g., mean, standard deviation) calculated in a predefined window as extracted features. Furthermore, classic machine learning models heavily depend on feature engineering and selection. Learning robust features is a challenging task owing to the complexity of physiologic processes and the nonlinear relationship between medical events. The engineering process of creating, analyzing, selecting, and evaluating appropriate features can be laborious and time-consuming.

In this paper, the proposed framework relies on deep neural networks to learn optimal features directly from the data itself without any human guidance, allowing the automatic discovery of latent data relationships that might otherwise be unknown. The proposed framework was compared against several commonly used classic machine learning models.

Recurrent Neural Network (RNN) and LSTM

Recurrent Neural Network (RNN) is one of the most extensively researched deep neural networks for handling temporal sequential data. LSTM is a type of RNN specifically designed to avoid gradient vanishing and exploding problems. RNN has been applied to many EHR applications (e.g., Doctor AI [21]) due to its memory maintenance mechanism [22] and parameter sharing scheme, which allow the model to capture long-range temporal dependency and to deal with sequences of varying length.

Recently, a lot of extensions have also been proposed to improve LSTM. For example, Che et al. focused on handling missing values and time irregularities [23]. DeepCare [24] modeled the effect of time irregularities through the activation of forget gates. The study also explored the confounding interactions between disease progression and interventions. Zhang et al. applied a LSTM based framework using two levels of imperfect yet informative labels to jointly learn the distinct patterns of septic shock. Our framework aims at exploring the effect of two additional components (i.e., CNN and static information) on improving LSTM. None of the aforementioned literature has explored the same structure.

Convolutional Neural Network (CNNs). Originally invented for computer vision, CNN has been shown to achieve remarkable results in many image/video related tasks [7]. Despite its huge success in vision, its applicability to EHRs is still unclear. Unlike image processing tasks, where convolution operation is applied over both width and height, most CNN applications using EHRs only apply convolution operations over the temporal dimension but not the attributes/feature dimensions. A study [26] closely related to our work compared one-side CNNs against both logistic regression and LSTM to predict the onset of multiple diseases. Their result showed both LSTM and CNN outperform logistic regression greatly, however, there is no clear winner between LSTM and CNN. Che et al. compared one-side CNN with Random Forest, however their results show CNN did not outperform Random

Forest.

Our framework is different in that applies LSTM on top of CNN to further capture its temporal dependencies, whereas the previous literature relies on fully connected neural network or temporal fusion techniques instead.

Adding Static Information. Esteban et al. [12] applies an independent fully connected neural network to process static information. The output is then combined with the hidden states of RNN at every time step. The models were applied to predict the outcome of patients with kidney failure. The results showed that the model's performance was improved greatly after static information was incorporated.

III. Septic Shock Challenges

There are two major challenges in septic shock prediction: the first is the subtle progression of septic shock, and the second is the lack of a well-established definition for septic shock.

The first challenge is due to the fact that 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 in white cell count and body temperature during the early stages of sepsis. However, predicting septic shock becomes increasingly challenging when the duration of the patient's visit spans a long period of time, as it is difficult for most sequential models to both memorize that much detail and consistently connect previous information to the present without significant loss. To overcome this challenge, septic shock progression is modeled using variants of Long-Short Term Memory (LSTM) networks because of its ability to memorize temporal dependencies over a long period.

The second challenge stems from the lack of a widely accepted definition for sepsis/septic shock and an ill-defined criteria for labeling different steps of its progression. This is evidenced by the fact that the definition of sepsis and septic shock are a point of disagreement between government agencies [28]. This disagreement is partly due to the subtle changes in symptoms characterizing sepsis such as: minor changes in mental status, white cell count, and blood glucose levels [29]. Moreover, sepsis has many subtypes with distinct disease processes. According to the PIRO sepsis staging framework [30], a sepsis subtype can be defined by predisposition (e.g., age, gender, past medical conditions, etc), infection type (e.g., pneumonia, urinary tract, etc), response (e.g., heart rate, respiratory rate, etc), and organ dysfunction (e.g., cardiovascular, metabolic, etc). Patients with a similar risk of mortality may express drastically different symptoms depending on their sepsis subtypes.

IV. Proposed Framework

Our proposed framework consists of three components: LSTM, CNN, and fully connected neural network (FC). A detailed description of each component is provided below.

A. LSTM

LSTM [6] is a type of recurrent neural network specifically designed to avoid the vanishing and exploding gradient problems. LSTM enables the network to maintain the previous information of hidden states as internal memory. Therefore, it is particularly suitable for tasks where long range temporal dependencies between events exist.

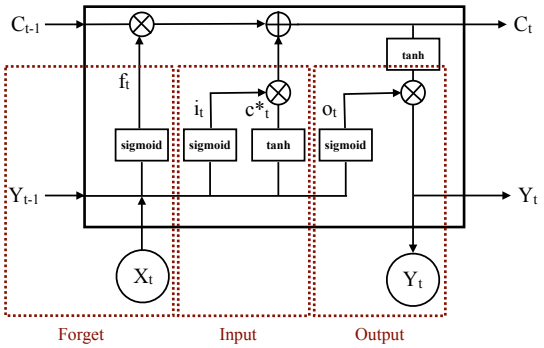

Fig. 1: A single LSTM block containing a forget, input and ouput gate

The architecture of a single LSTM block is shown in Figure 1. It consists of a memory cell state denoted by C_t and the following three gates: Forget gate $f_t \in [0, 1]$, Input gate $i_t \in [0, 1]$, and Output gate $o_t \in [0, 1]$. These three gates interact with each other to control the flow of information. During training, the network learns what to memorize and when to allow reading/writing in order to minimize the classification errors. More specifically, the Forget gate determines what information from the previous memory cell state is expired and should be removed; the Input gate selects information from the *candidate* memory cell state C_t^* to update the cell state; the Output gate filters the information from the memory cell so that the model only consider information relevant to the prediction task. The value of each gate is computed as follows, where $W_{[i,f,C,o]}$ are the weight matrices and $b_{[i,f,C,o]}$ are the bias vectors:

$$\begin{aligned} i_t &= sigmoid(W_i \cdot [y_{t-1}, X_t] + b_i) \\ f_t &= sigmoid(W_f \cdot [y_{t-1}, X_t] + b_f) \\ C_t^* &= \tanh(W_C \cdot [y_{t-1}, X_t] + b_c) \\ o_t &= sigmoid(W_o \cdot [y_{t-1}, X_t] + b_o) \end{aligned} \tag{1}$$

The memory cell value C_t and output label y_t from the LSTM block are computed using the following formulas:

$$\begin{aligned} C_t &= C_{t-1} \cdot f_t + C_t^* \cdot i_t \\ y_t &= O_t * tanh(C_t) \end{aligned} \tag{2}$$

For our task, only the output label at the last time step T is obtained to make the prediction.

B. Convolutional Neural Network (CNN)

Our framework employs a CNN component as shown in Figure 2 to extract local, temporal-invariant dependency. The input of CNN is dynamic information represented as temporal sequences; the output of CNN is compact state representations that are further fed into LSTM in the next step.

Our CNN component consists of three types of layers: the Convolutional Layer, the Rectified Linear Unit (ReLU) Layer, and the Pooling Layer. The Convolutional Layer employs a set of convolutional filters that look for specific patterns, the ReLU layer introduces non-linearity and it is efficient to compute as it has constant derivatives. The pooling layer reduces the spatial/temporal span of the extracted features.

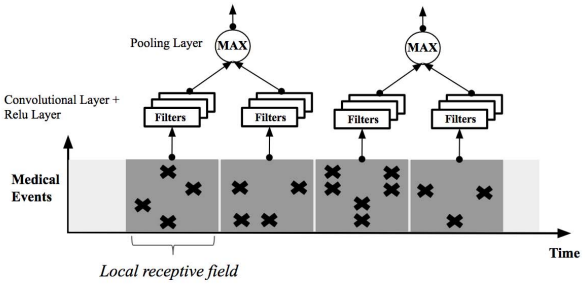

Fig. 2: Convolutional Neural Network (CNN) extracts state representation from dynamic information

The convolutional Layer employs a set of convolutional filters. The filters always span the entire feature dimension, but are connected to only a small temporal region of the data. They carry out a set of computations to determine whether specific patterns shows up. Then the filters traverse along the temporal direction of the sequence, producing a sequence of outputs indicating where different patterns occur. Note that when CNN is applied to image tasks, the filters were allowed to travel through both width and height dimensions. However, when it is applied to EHRs, the convolutional filters move *only in the temporal dimension*.

This demonstrated CNN architecture has the following distinguishing properties: local receptive fields, shared weights, and temporal sampling [31]. When dealing with temporal sequences, it is impractical to extract features by connecting neurons to all time steps. Therefore, CNN employs a set of filters which only connects its neurons to a local temporal region. The temporal extent of the connectivity is a hyperparameter called *local receptive field*. Figure 2 shows an example of a convolutional layer with a local receptive field of size 3. Parameter sharing is employed to limit the number of parameters. It is based on the reasonable assumption that if a pattern exists at some temporal location, then it can also exist in other temporal location. More specifically, the convolutional filters traverse from one location to the next, carrying out a set of computations to look for some specific patterns. Both local receptive fields and shared weights allow the model to learn *local* and *time-invariant* features effectively. Finally, the pooling layer performs temporal sampling by taking an average or using the min/max operation, thereby reducing the model's sensitivity to shifts and distortion. Note that the CNN is capable of learning its own feature extractors automatically through backpropagation. The state representations make the local characteristics explicit so that LSTM can recognize the temporal patterns more easily.

C. Static Information

Two different approaches were implemented to incorporate static information into our framework. The first approach (Figure 3) is proposed by previous literature [12]. The static information is first processed by FC. The output is then concatenated with the hidden states of LSTM at *every* time step. We denote this approach as *Static-repeat*.

Our proposed approach (Figure 4) also employs FC to process static information. However, the output is concatenated with the output of LSTM only in the *last or final* time step. The concatenated vector is used to make the final septic shock prediction. We denote the second approach as *Static-last*.

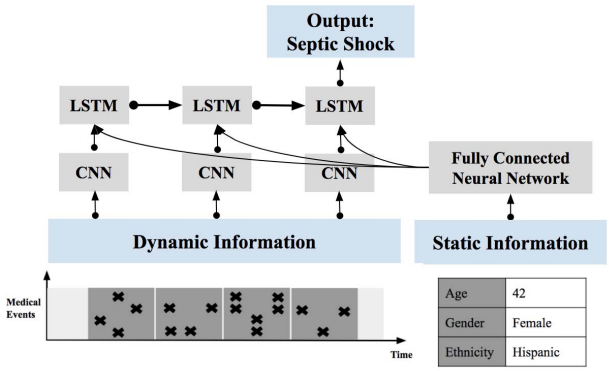

Fig. 3: LSTM+CNN+Static-repeat: the output of FC is concatenated with the hidden state of LSTM at every time step

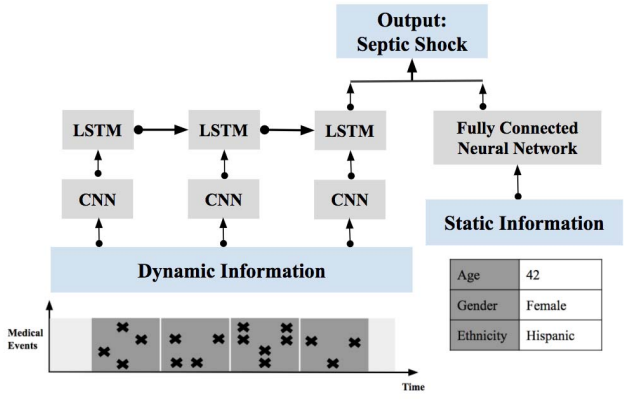

Fig. 4: LSTM+CNN+Static-last: the output of FC is concatenated with the output of LSTM at last step

V. DATASET

Data Descriptions This study used de-ientified EHRs obtained from Christiana Care Health System (CCHS), Newark, DE. The CCHS dataset includes retrospectively collected EHR data for adult patients (age $\geqslant$ 18 years) hospitalized within CCHS from July 2013 to December 2015, corresponding to 119,968 unique patients and 210,289 hospitalizations.

The EHRs contain both *static* and *dynamic* information. *Static* information contains patient background information collected once per visit. Our study uses 35 *static* variables. *Dynamic* information is collected multiple times at irregular intervals during the patients' entire hospitalization and has a time stamp associated with each record; hence, *dynamic* information can be expressed as a time sequence. Our study includes 43 dynamic variables. The list of static and dynamic variables is shown in Figure 5.

Static Information
• Gender • Age • Ethnicity • Race • Flu season • 30 comorbid conditions (e.g.,Diabetes)
Dynamic Information
• 6 vitals (e.g., heart rate) • 4 treatment (e.g., antibiotic) • 11 laboratory and assessment results (e.g., lactate, platelet) • 18 culture results (e.g., PCR influenza culture) • 4 other categories (e.g., oxygen source)

Fig. 5: Static and dynamic variables

Study Population The *study population* are patients with *suspected infection* which was identified by the presence of any type of antibiotic, antiviral, or antifungal administration, or a positive test result of Point of Care Rapid (PCR). Note that the study population, the aforementioned rules for identifying suspected infection, and labeling criteria for septic shock in the following, were all determined by the two leading clinicians with extensive experience on this subject from Mayo Clinic and Cristina Care Health System.

Ground Truth Labeling for Septic Shock: Supervised models depend heavily on the accurate label of the training dataset. However, acquiring the true label (i.e., septic shock and non septic shock) can be challenging. Diagnosis codes, such as International Classification of Diseases, Ninth Revision (ICD-9), are widely used. However, solely relying on ICD-9 can be problematic as it has been proven to have limited reliability due to the fact that its coding practice is used mainly for administrative and billing purposes. Indeed, it has been widely argued that ICD-9 codes cannot be used for establishing reliable gold standards for various clinical conditions [32]. More importantly, ICD-9 cannot tell when septic shock occurs, which is essential for our task. On the basis of the Third International Consensus Definitions for Sepsis and Septic Shock [33], our domain experts identified septic shock if any of the following three clinical rules is met:

- Persistent hypotension as shown through two consecutive readings (>=30 minutes apart)
  - Systolic Blood Pressure (SBP) $<$ 90 mmHg
  - Mean arterial pressure (MAP) $<$ 65 mmHg
- A decrease in SBP >= 40mmHg within an 8-hour period
- Any vasopressor administration

When applying both ICD-9 and our clinical rules, we identified 1,869 shock positive visits and 23,901 negative visits. Given the imbalanced ratio of positive and negative shock visits, we further conducted a stratified random sampling on shock negative visits while keeping the same underlying distribution of age, gender, ethnicity, length of stay and the number of records in both positive and negative visits. As a result, the final dataset has 3,738 visit s (1,869 positives and 1,869 negatives) and 145,421 events. Figure 6 shows the septic shock onset time as x-axis and the percentage of the shock patients as y-axis.

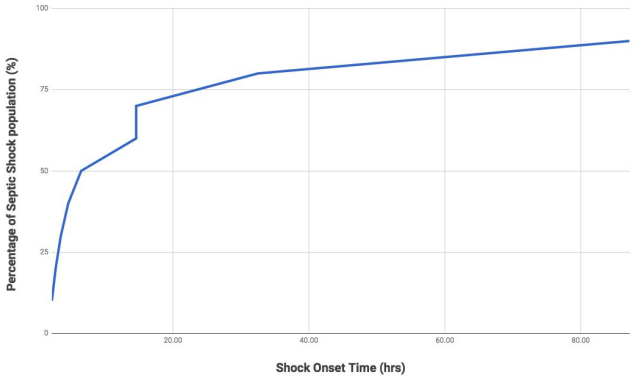

Fig. 6: The percentage of septic shock population and their septic shock onset time

VI. EXPERIMENTS

We first compared among the 6 deep neural network models:

- The original LSTM: LSTM-Origin
- Adding static information only: LSTM+Static-repeat and LSTM+Static-last
- Adding CNN only: LSTM+CNN
- Adding both CNN and static information: LSTM+CNN+Static-repeat, LSTM+CNN+Static-last

The best model was selected and compared against the 6 machine learning baselines: Logistic Regression, Gaussian naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest and Multi-layer Perceptron (MAP).

A. Two Different Early Diagnosis Tasks

All the models were tested using two different early diagnosis tasks: the visit level early diagnosis (left aligned) and the event level early prediction (right aligned).

For the visit level early diagnosis task (Figure 7), the *first* n hours of the patients' EHRs were made available to the models. The goal is to predict whether the patient will develop septic shock at any subsequent point during the visit. To carry out this task, the patients' sequences were left aligned, i.e., they were aligned by the start of their visits. Only records within the first n hours of patients' visit were used for training and testing. This n-hour window is denoted as *observation window* and is represented as a shaded area in Figure 7.

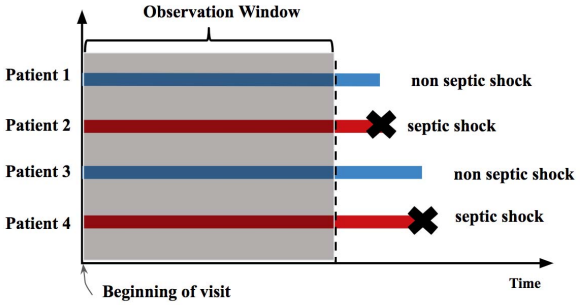

Fig. 7: Left align: predict whether a patient will have septic shock at the end of the visit using records in the first n-hour window

For the event level early prediction task, our goal is to predict whether the patient will develop septic shock m hours later. To carry out this task, we *right aligned* all sequences by their end point. For septic shock patients, the end point is the onset time of septic shock, whereas for non-septic shock patients, the end point is the end of sequences. The EHRs within the m-hour window leading up to endpoint were omitted. This m-hour window is denoted as hold-off window. Only the records before the hold-off window (shaded area) were used to predict what would happen m hours later.

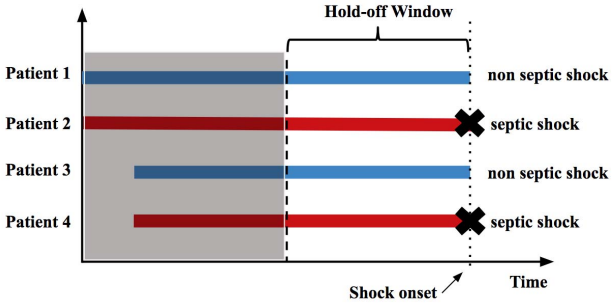

Fig. 8: Right align: predict whether a patient will have septic shock m hours later

B. Implementation

For classic machine leanring baselines, 174 features were extracted as these models do not handle time series directly. Motivated by previous literature [20], the mean, max, min, median, and standard deviation were calculated as features for numeric variables. For categorical variables, we counted how many times the variable was collected and how many times a given value was observed. That static information was also included. The 6 machine learning baselines were all implemented using the scikit-learn package in Python.

The deep neural network based approaches require us to address missing values and time irregularities in time series. To address missing values, we adopted a zero imputation and missing indicator strategy originally proposed by Lipton et al. [34]. An indicator variable M_t was introduced for every X_t, where $M_t = 1$ if X_t is missing and 0 otherwise. Then zero was imputed in all missing entries. The models are able to extract the missing patterns from those indicators. To address time irregularities, we hypothesized that the frequency of clinical events imply the stability of a patient's health: patients with more severe clinical conditions normally receive more intensive treatment, which can be reflected by the time intervals between two consecutive medical events. Therefore, a new feature $Time_t$ capturing the time interval between consecutive medical events was introduced. All the deep neural network based models were implemented in Keras with Tensorflow as the backend engine. For LSTM layer, we experimented 20, 40, 60, and 80 units and the best one was chosen. For CNN, 100 filters with filter size = 4 and pooling size = 3 were used. The FC has 20 hidden units. Binary Cross-entropy was applied as loss function and Adam optimizer was used for optimization. Early stopping were applied to avoid over fitting. All models were evaluated using 3-fold cross validation.

C. Evaluation Metrics

AUC, F1 Score, accuracy, recall, and precision are the most commonly used evaluation metrics. Accuracy tells the fractions of patients whose labels were correctly identified. Recall tells us what proportion of patients that actually had septic shock were correctly diagnosed. Precision tells us what proportion of patients who were diagnosed as having septic shock actually had septic shock. Finally, AUC calculates the tradeoff between recall and specificity. F1 Score is the harmonic mean of Precision and Recall that sets their tradeoff. Therefore, in the following we will mainly use F1-score and AUC to compare different models.

VII. Results

In the following sections, we present the results for both experimental settings.

A. Septic Shock visit level early diagnosis (left align)

Figure 9 shows the AUC and F1 Score for visit level early diagnosis while varying the observation window from 3 hours to 24 hours by 1 hour increments. For all six deep neural network based models, both evaluation metrics improve over time because more information became available as the observation window became larger. Among these six models, LSTM+CNN+Static-last achieved the best AUC at all points in time, while generally having the highest F1 Score (exceptions are 6 hr, 15 hr and 21 hr).

Next, we explored how different components would impact the model's performance. Table I shows the average

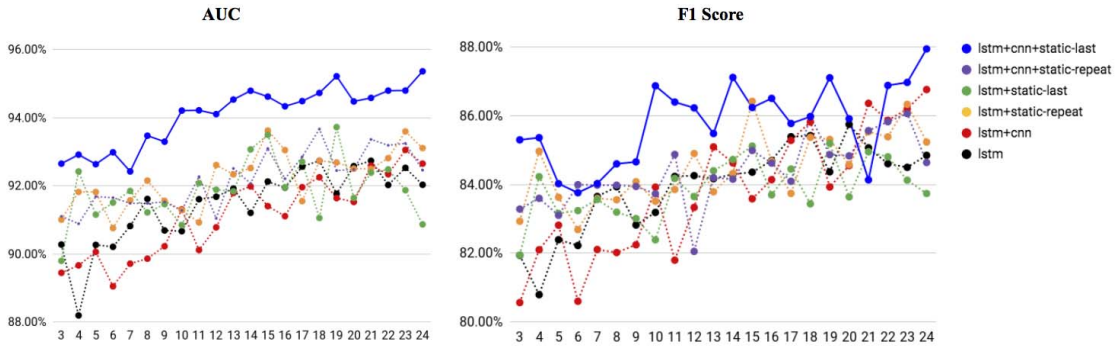

Fig. 9: AUC and F1 Score for septic shock early diagnosis while varying the size of observation window from 2 hr to 24 hr

TABLE I: Average AUC and F1 Score for septic shock visit level early diagnosis

	AUC ± std	**F1 Score ± std**
LSTM-Origin	91.46 ± 1.10	83.95 ± 1.25
LSTM+CNN	91.12 ± 1.18	83.81 ± 1.85
LSTM+Static-repeat	92.21 ± 0.83	84.47 ± 1.03
LSTM+Static-last	91.88 ± 0.91	83.86 ± 0.86
LSTM+CNN+Static-repeat	92.18 ± 0.82	84.38 ± 0.98
LSTM+CNN+Static-last	**94.08** ± 0.89	**85.79** ± 1.19

Notes: std stands for standard deviation.
Best metrics are highlighted in bold.

TABLE II: Results on septic shock visit level early diagnosis using records up to 12 hours

Model	**F1 Score**	**AUC**	**Accuracy**	**Recall**	**Precision**
1 Logistic Regression	71.62	79.85	72.35	69.77	73.56
2 Gaussian Naieve Bayes	68.62	71.47	60.64	**86.07**	57.05
3 SVM	71.27	77.07	70.31	73.63	69.05
4 Decision Tree	64.10	64.62	64.52	63.34	64.87
5 Random Forest	72.15	79.77	72.80	70.47	73.92
6 MAP	56.45	48.96	49.20	65.86	49.40
7 LSTM-Origin	84.26	91.68	84.59	82.48	86.12
8 LSTM+CNN+Static-last	**86.23**	**94.11**	**86.58**	84.08	**88.49**

TABLE III: Results on septic shock event level early prediction with hold-off window size from 2 hr to 4.5 hr (left) and from 5 hr to 24 hr (right)

	from 2 hr to 4.5 hr		from 5 hr to 24 hr	
	AUC ± std	**F1 Score ± std**	**AUC ± std**	**F1 Score ± std**
LSTM-Origin	82.76 ± 2.29	73.18 ± 2.72	74.35 ± 2.54	65.39 ± 2.75
LSTM+CNN	81.67 ± 2.22	73.61 ± 1.72	**77.17** ± 1.59	**69.09** ± 2.05
LSTM+Static-repeat	82.25 ± 3.32	73.31 ± 3.77	69.13 ± 6.15	63.05 ± 7.16
LSTM+Static-last	78.15 ± 7.18	70.95 ± 3.47	56.35 ± 6.69	51.81 ± 13.32
LSTM+CNN+Static-repeat	67.06 ± 11.06	57.63 ± 12.53	56.47 ± 6.54	51.22 ± 9.82
LSTM+CNN+Static-last	**85.17** ± 1.97	**76.74** ± 2.30	72.65 ± 7.27	64.14 ± 7.13

AUC and F1 Score, as well as their standard deviation, as the size of the observation window increased from 3h to 24h. Comparing LSTM-Origin and LSTM+CNN, it is evident that adding the CNN component alone would not improve LSTM's performance. Comparing both LSTM+Static-repeat and LSTM+Static-last against LSTM-Origin, it was found that adding static information repeatedly at every time step improves LSTM, whereas adding static information only at the last step does not improve it. The CNN component was then introduce to models with static information, and it was found that the combination of CNN and Static-repeat does not improve LSTM, while the combination of CNN and Static-last greatly improved LSTM. This is evident in the fact that LSTM+CNN+Static-last achieves the highest average AUC=94.08% and F1 Score=85.79% across all models.

Since LSTM+CNN+Static-last achieved the best results, it was further compared against six classic machine learning baselines and the deep neural network baseline LSMT-Origin. The observation window size was set to be 12 hr. Note that the same patterns were found across 3 hr to 24 hr. As table II shows both deep neural network based approaches LSTM-origin and LSTM+CNN+Static-last out-

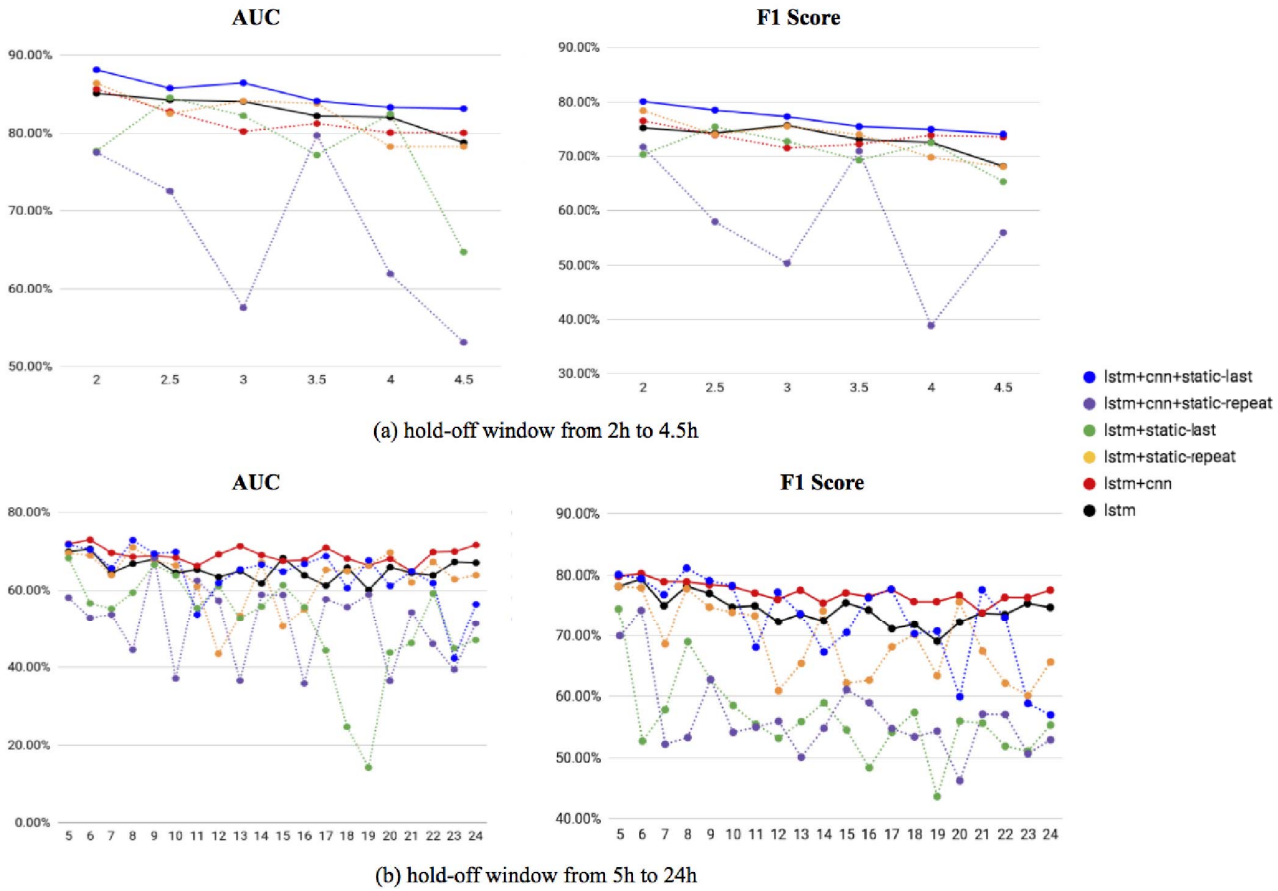

Fig. 10: AUC and F1 Score for septic shock event level early prediction while varying the size of hold-off window from 2h to 4.5 h (above) and from 5h to 24 h (below)

perform the six classic machine learning baselines (row 1-6) in all five measures except recall. Gaussian Naive Bayes achieves the best recall=86.67% while both LSTM-Origin and LSTM+CNN+Static-last have reasonable high recall: 82.48% and 84.05% respectively. When comparing LSTM-Origin with LSTM+CNN+Static-last, it can be seen that the latter outperforms the former in every measure.

In conclusion, for left align visit level early diagnosis, our proposed framework LSTM+CNN+Static-last achieves the best results. Thus, the combination of the CNN component and static information (via static-last) benefits the visit-level diagnosis.

B. Septic Shock Event Level Early Prediction (Right Align)

Figure 10 shows the results for septic shock event level early prediction while varying the size of the hold-off window from 2 hr to 24 hr. Our results showed that different patterns exist for $<$ 5 hours and $\geq$ 5 hours. Therefore, we present the performance using two figures.

The task becomes more challenging because less records are available as the size of hold-off window increases. Figure 10(a) presents the AUC and F1 Score when hold-off window ranges from 2 hr to 4.5 hr using a 0.5-hour increment. It can be seen that LSTM+CNN+Static-last consistently achieves the best AUC and F1 Score. Figure 10(b) shows the AUC and F1 Score when hold-off window ranges from 5 hr to 24 hr using a 1-hour increment. From this figure, we can see LSTM+CNN has the best AUC and F1 Score the majority of time. More importantly, LSTM+CNN has the most stable behavior when compared to all other models.

Table III shows the average AUC and F1 Score as well as their standard deviation. The second column shows the results when the hold-off window is between 2 to 4.5 hr; whereas the third column shows the results when the hold-off window is between 5 to 24 hr. Overall, LSTM+CNN+Static-last achieves the best AUC and F1 Score when the hold-off window is between 2 to 4.5 hr. LSTM+CNN achieves the best metrics when the hold-off window is between 5 to 24 hr.

Next, we explore how each component affects the model. When hold-off window is between 2 to 4.5 hr, neither adding CNN alone nor adding static information alone helps the model. When the combination of CNN and static information is considered, LSTM+CNN+Static-repeat also does not benefit the model. Nonetheless, LSTM+CNN+Static-last improves the LSTM significantly. When the hold-off window is between 5

TABLE IV: Results on septic shock event level early prediction (above: hold-off window = 3 hr; below: hold-off window = 6 hr)

Model	F1 Score	AUC	Accuracy	Recall	Precision
1 Logistic Regression	68.28	74.20	68.89	67.00	69.64
2 Gaussian Naieve Bayes	67.05	68.74	62.76	75.79	60.12
3 SVM	64.75	70.47	63.52	67.00	62.64
4 Decision Tree	59.50	59.66	59.72	59.17	59.82
5 Random Forest	69.10	74.04	69.22	68.84	69.37
6 MAP	68.28	74.30	58.58	67.43	69.15
7 LSTM-Origin	75.71	84.08	76.76	72.42	**79.31**
8 LSTM+CNN+Static-last	**77.31**	**86.47**	**77.47**	**76.76**	77.86

Model	F1 Score	AUC	Accuracy	Recall	Precision
1 Logistic Regression	66.93	73.82	67.63	65.50	68.41
2 Gaussian Naieve Bayes	59.32	67.09	62.69	54.41	65.21
3 SVM	67.18	72.84	67.48	66.57	67.80
4 Decision Tree	64.11	69.16	63.75	64.74	63.49
5 Random Forest	56.12	56.38	56.38	55.78	56.46
6 MAP	67.14	73.83	68.09	65.20	69.19
7 LSTM-Origin	70.65	79.35	71.66	68.24	73.25
8 LSTM+CNN	**73.00**	**80.25**	**72.34**	**74.77**	**71.30**

to 24 hr, adding CNN alone benefits LSTM greatly, whereas adding static information alone, or adding both CNN and static information does not help.

Finally, we compared the best models with 6 machine learning baselines and LSTM-Origin. Table IV(above) shows the results of event level early prediction with a 3-hour hold-off window. Both LSTM-Origin and LSTM+CNN+Static-last outperform classic machine learning models. LSTM+CNN+Static-last achieves the best results in all measures except precision. Table IV(below) shows the results of the same task with a 6-hour hold-off window. The results show that both LSTM+Origin and LSTM+CNN outperform machine learning baselines. LSTM+CNN has the best results for all measures.

To summarize, for septic shock event level early prediction, when the hold-off window is less than 4.5h, adding both CNN and static information improves the performance; when the hold-off window is more than 4.5 hr, adding CNN alone achieves better performance.

VIII. Conclusions

In this work, we presented a generic framework based on LSTM. We further investigated how the model would perform if CNN and static information are incorporated. The proposed framework was tested on two experimental settings: visit level early diagnosis (left align) and event level early prediction (right align). We conclude that it is not always necessary to add all components. More specifically, for visit level early diagnosis, adding both CNN and static information would benefit the performance. Most notably, the propose framework LSTM+CNN+Static-las achieves an AUC above 92.00% and F1 Score above 85.00% when only the first 3 hours of the EHRs were used. For event level early prediction task, the results differ according to the size of hold-off window. When the hold-off window is less than 4.5 hr, the proposed framework LSTM+CNN+Static-last achieves the best results. More importantly, LSTM+CNN+Static-last can predict whether a patient will have septic shock 4 hours later with AUC above 80.00% and F1 Score above 75.00%. However, when the hold-off window ranges from 5 hr to 24 hr, adding CNN alone would have better and more stable results.

For future work, we would like to further explore why LSTM+CNN+Static-last does not help for event level early prediction when the hold-off window $\geq$ 4.5 hours, which is the most challenging task. Second, we would like to explore the model's behavior for different patient subtypes. Finally, we want to test the models on different datasets (e.g., MIMIC) to see if the same results are achieved.

Acknowledgment

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

References

[1] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annual review of public health*, vol. 36, pp. 345–359, 2015.

[2] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, "Multistate markov models for disease progression with classification error," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 2, pp. 193–209, 2003.

[3] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 85–94.

[4] S. Ghosh, J. Li, L. Cao, and K. Ramamohanarao, "Septic shock prediction for icu patients via coupled hmm walking on sequential contrast patterns," *Journal of biomedical informatics*, vol. 66, pp. 19–31, 2017.

[5] K. Orphanou, A. Stassopoulou, and E. Keravnou, "Temporal abstraction and temporal bayesian networks in clinical domains: A survey," *Artificial intelligence in medicine*, vol. 60, no. 3, pp. 133–149, 2014.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[10] H. R. Wong, C. J. Lindsell, V. Pettilä, N. J. Meyer, S. A. Thair, S. Karlsson, J. A. Russell, C. D. Fjell, J. H. Boyd, E. Ruokonen *et al.*, "A multibiomarker-based outcome risk stratification model for adult septic shock," *Critical care medicine*, vol. 42, no. 4, p. 781, 2014.

[11] J. Cohen, J.-L. Vincent, N. K. Adhikari, F. R. Machado, D. C. Angus, T. Calandra, K. Jaton, S. Giulieri, J. Delaloye, S. Opal *et al.*, "Sepsis: a roadmap for future research," *The Lancet infectious diseases*, vol. 15, no. 5, pp. 581–614, 2015.

[12] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," in *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE, 2016, pp. 93–101.

[13] R. P. Dellinger, J. M. Carlet, H. Masur, H. Gerlach, T. Calandra, J. Cohen, J. Gea-Banacloche, D. Keh, J. C. Marshall, M. M. Parker *et al.*, "Surviving sepsis campaign guidelines for management of severe sepsis and septic shock," *Intensive care medicine*, vol. 30, no. 4, pp. 536–555, 2004.

[14] R. Frost, H. Newsham, S. Parmar, and A. Gonzalez-Ruiz, "Impact of delayed antimicrobial therapy in septic itu patients," *Critical Care*, vol. 14, no. S2, p. P20, 2010.

[15] J. M. Bohnen, R. A. Mustard, S. E. Oxholm, and B. D. Schouten, "Apache ii score and abdominal sepsis: a prospective study," *Archives of Surgery*, vol. 123, no. 2, pp. 225–229, 1988.

[16] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (saps ii) based on a european/north american multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.

[17] e. a. Vincent, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive care medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[18] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[19] e. a. Rivers, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *New England Journal of Medicine*, vol. 345, no. 19, pp. 1368–1377, 2001.

[20] D. Shavdia, "Septic shock: Providing early warnings through multivariate logistic regression models," Ph.D. dissertation, Massachusetts Institute of Technology, 2007.

[21] E. Choi, M. T. Bahadori, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *arXiv preprint arXiv:1511.05942*, 2015.

[22] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[23] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *arXiv preprint arXiv:1606.01865*, 2016.

[24] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2016, pp. 30–41.

[25] Y. Zhang, C. Lin, M. Chi, J. Ivy, M. Capan, and J. Huddleston, "Lstm for septic shock: Adding unreliable labels to reliable predictions," pp. 1233–1242, 12 2017.

[26] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Machine Learning for Healthcare Conference*, 2016, pp. 73–100.

[27] Z. Che, Y. Cheng, Z. Sun, and Y. Liu, "Exploiting convolutional neural network for risk prediction with medical feature embedding," *arXiv preprint arXiv:1701.07474*, 2017.

[28] e. a. Seymour, "Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 762–774, 2016.

[29] R. S. Hotchkiss and I. E. Karl, "The pathophysiology and treatment of sepsis," *New England Journal of Medicine*, vol. 348, no. 2, pp. 138–150, 2003.

[30] S. Rathour, S. Kumar, V. Hadda, A. Bhalla, N. Sharma, and S. Varma, "Piro concept: staging of sepsis," *Journal of postgraduate medicine*, vol. 61, no. 4, p. 235, 2015.

[31] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[32] K. Giuliano, "Physiological monitoring for critically ill-patients: testing a predictive model for the early detection of sepsis," *AJCC*, vol. 16, no. 2, March 2007.

[33] e. a. Singer, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[34] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.