Sepsis Prediction: An Attention-Based Interpretable Approach

Kourosh T. Baghaei
Computer Science and Engineering
Mississippi State University
Starkville MS, USA
kt1414@msstate.edu

Shahram Rahimi
Computer Science and Engineering
Mississippi State University
Starkville MS, USA
rahimi@cse.msstate.edu

Abstract—**Sepsis is the leading cause of death in ICUs and a very costly medical phenomena. The earlier it is predicted, the less inpatient mortality and the less the length of ICU stay, thus a major cut in medical expenses. Although the current deep learning models are able to make predictions about the possibility of sepsis in the ICU, they still lack the ability to reveal the major factors that lead to the outcomes of the predictions. In this paper, we have explored the use of an attention-based model in prediction of sepsis which provides more details on the amount of contribution of each of the medical measurements to the final prediction. This would help health care providers to improve their procedures to reduce sepsis related mortality rate.**

I. Introduction

Sepsis is defined as a syndrome of abnormal biochemical, physiologic, and pathologic status caused by infection [1]. As the leading cause of death in ICUs all over the world, which accounts for over $23 billion of medical expenses in the United States, it is a major medical concern [2]. This is while with early detection of sepsis and proper intervention, up to 80 percent of in-hospital sepsis-related deaths can be prevented [3].

To help physicians to diagnose sepsis, several medical scoring systems have been proposed as guidelines in the ICU, such as SOFA [4], APACHE II [5], and MEWS [6]. Moreover, with the advances in technology and AI, more comprehensive and efficient approaches to data analysis for this purpose have been made possible. The Electronic Health Records (EHR) alongside statistical and deep learning approaches provide new insights from the physiological and biochemical measurements and diagnoses that help physicians in predicting medical outcomes and intervening accordingly [7].

Despite the promising predictive capabilities of the deep models [8]–[13] in terms of accuracy, precision, AUROC, etc., due to their non-linearity and high complexity, they act as *blackbox* that provide the user with sole probability values [14]. Whereas, in the health care context, understanding the causes that lead to a certain prediction can make a difference in a life threatening situation.

In this research work, we implemented an attention-based deep model for the prediction of sepsis that visualizes the extent to which a medical parameter affects the outcome of the prediction. In section (II) a review of the related works is provided. In section (III), the data set used in our study is described. We then explain our methodology in section (IV), and discuss the results in section (V). Finally conclusion and future works are presented in the last section.

II. Related Work

There has been efforts to define some standards and definitions that could help medical doctors identify sepsis during an ICU stay [4]–[6]. However, the AI based predictive deep models look more and more promising and are attracting more researchers every day. Based on the format of input data, we can group the approaches into two categories. In one approach, ICD9[1] codes, that are actually procedures billed during each visit of patients to health centers, are used in order to predict the medical outcomes of patients [15]–[19]. In the other major approach, instead of using ICD9 codes, electronic health records (EHR) such as heart rate, respiratory rate, systolic blood pressure, etc. are used as multivariate time series [3], [7]–[10], [13], [20], [21]. Additionally, as in [11], there exist few approaches in which both of these data types are utilized.

In spite of the fact that both approaches yield high quality results in prediction, they have a few differences that should be considered: First, the former approach highly depends on the billed procedures for each patient's visit without considering any further details about the corresponding visit and patient's health status. On the other hand, in the latter approach, more detailed information about patient's health status is provided. Thus making decisions/predictions based on the patterns in the EHR would be more precise and realistic.

Second, the research papers that follow the former approach tend to map the ICD9 labels to the machine translation problems [22] for making predictions. Hence, they mostly make use of variations of Recurrent Neural Networks (RNN) such as LSTM and GRU [23]. However, in studies where inputs are multivariate time series, other architectures are proposed such as in [3] or in [7] that a CNN-based model is applied to this problem. Additionally, a model called Artificial Intelligence Sepsis Expert has been proposed [21] which is based on Weibull-Cox proportional hazards model [24]. Another statistical model called Insight was proposed in [13]

[1] International Classification of Diseases, 9th Revision

978-1-5386-1728-1/19/$31.00 ©2019 IEEE

for sepsis prediction. Furthermore, [8]–[11] have used RNNs for multivariate time series.

Although both of the aforementioned approaches result in high quality predictions, they mostly tend to perform as a *black-box* which lacks the ability to explain what has led to a certain result. This is while naturally the physicians would make more reliable preventive decisions, if they are aware of such information. The issue of *interpretability* is addressed for ICD9 codes in [16]–[19]; however, the drawbacks of using ICD9 codes still remain. On the other hand, for the EHR data, interpretable models are proposed which do not have the drawbacks of ICD9 [3], [8], [21].

In this study, we explore the use of an attention based RNN with GRU cells in order to make visually interpretable predictions, addressing the shortcomings of ICD9 based approaches. Although this is a classification problem, classifying a patient as septic or non-septic, we simply use the word *prediction* throughout this paper.

III. Data Set

For this research project, we have used MIMIC-III data set [25], which is a freely accessible database developed by the Laboratory of Computational Physiology at MIT. This data set consists of vital signs, laboratory measurements, diagnostic codes, procedure codes, observations and notes provided by medical staff and so on for over 53000 adult patients. In this work, we have used only a subset of this database which is explained in the next section.

A. Gold Standard

The general definition of sepsis can be stated as the life-threatening organ failure in response to inflammatory response to infection. Technically, in order for a patient to be considered septic, a set of certain parameters should meet the criteria of the Third International Consensus Definitions for Sepsis and Septic Shock [1]. We have utilized [2], [26] to produce our cohort of study and label patients as septic and non-septic based on Sepsis 3 Definition [1]. In total there are 11,791 patients in the cohort studied in [2]. However, we did not consider the ones that were either too sparse or too short, which resulted in total number of 11,700 patients.

B. Feature Selection and Preprocessing

The definition of Sepsis 3 [1] provides a flowchart which uses SOFA [4] and qSOFA [27] scores alongside laboratory measurements for determining a patient as septic. We chose these features for our study, in total 39 Features, including 7 vital signs and 32 laboratory and output values. We aggregated data to one-hour bins. The missing values were filled with forward filling, and we interpolated the intermediate missing values for the training set. As in [12], we then discretized each of these continuous values into bins as follows:

- Lower than Normal
- Normal
- Higher than Normal

We used [28] as reference of normal values and considered the age and gender in discretization process. We also consulted medical experts for a few of the reference values. Where no value was present for any laboratory measurements at all, as in [11] we assumed that medical doctors deemed the parameter irrelevant to the patients' status. Otherwise, they would have measured it. Thereupon, we considered them as normal. The lengths of multi-variate sequences of patients are different across our cohort which is the case for the real world settings.

IV. Methodology

In this study we use RNNs in order to make predictions given multi-variate time series of patients' data. Among the variations of the RNNs, we use GRU [22] in our implementations and throughout this paper, and we simply refer to it by RNN. In machine translation problems, the sequences of the words contribute to the final output of the RNN, and therefore, to the outcome of the translation [22]. Likewise, we have assumed the sequence of the different statuses of the features contribute to the final outcome of the patients' status. Thus, we use *attention* to calculate the context for the variables as in [16], [19] in order to understand what parameters have contributed most to the final outcome. In the following sections, we first explain the inputs and the time series then the model is explained, and finally, we explain how the results are actually interpreted.

A. Problem Setup

For each patient i we have a multi-variate sequence P_i = $[x_1, x_2, x_3, ..., x_n]$ where $x_t \in \{0,1\}^r and 1 \leq t \leq n$. x_t is a vector representing status of different physiological and laboratory variables at time-step t of the sequence, and r is the total number of possible states of features that were explained in section (III-B). For each of the aforementioned states, the corresponding element of vector x_t is one.

B. Prediction Model

The high level architecture of our model is depicted in Fig. 1. Essentially, we have employed the prediction model proposed in [16] as the core of our model as is explained below. In order to consider the influence at the time-step level and the variable level (individual elements of x_t), a linear embedding is used as follows:

$$v_t = W_{emb} x_t \tag{1}$$

The vector v_t is the linear embedding of the input vector x_t, m is the size of the embedding dimension and $W_{emb} \in m \times r$ represents the embedding matrix.

We employ two sets of weights, one for calculating the attention at time-step level and the other for calculating the attention at variable level. The set of scalars $\alpha_1, \alpha_2, ..., \alpha_n$ hold the amount of influence of time-step t on the final outcome and vectors $\beta_1, \beta_2, ..., \beta_n$ represent the influence of each coordinate of embedding vector $v_{1,1}, v_{1,2}, v_{1,m}, ..., v_{n,1}, v_{n,2}, ..., v_{n,m}$.

Unlike [16], we do not use unidirectional RNNs. Rather, we employ bidirectional RNNs as in [17]–[19]. Bidirectional RNNs are becoming a point of interest in time series analysis. With their ability to produce two different sets of hidden states, from backward and forward iterating, they tend to outperform their unidirectional counterparts [19]. In order to generate $\alpha' s$ and $\beta' s$ separately, we employ two biderctional RNNs, BiRNN$_\alpha$ and BiRNN$_\beta$. The hidden states are calculated for both directions for each of the BiRNNs. For each t, $1 \leq t \leq n$ we calculate hidden states of BiRNNs as follows:

$$\begin{aligned} [g_t^f; g_t^b] &= BiRNN_\alpha(v_t) \\ g_t &= [g_t^f; g_t^b] \\ e_t &= w_\alpha^T g_t + b_\alpha \\ [h_t^f; h_t^b] &= BiRNN_\beta(v_t) \\ h_t &= [h_t^f; h_t^b] \end{aligned} \tag{2}$$

We then calculate $\alpha' s$:

$$\alpha_1, \alpha_2, ..., \alpha_n = softmax(e_1, e_2, ..., e_n) \tag{3}$$

For $t = 1$ to n the variable-level attentions are calculated as:

$$\beta_t = tanh(W_\beta h_t + b_\beta) \tag{4}$$

We can calculate the context vector c_t for the time-step t using the equation below:

$$c_t = \sum_{j=1}^{t} \alpha_j \beta_j \odot v_j \tag{5}$$

In the above equation, $\odot$ represents an element-wise multiplication. Using context vector $c_t \in \mathbb{R}^m$, we can predict the true label $y_t \in \{0, 1\}^2$ as follows:

$$\hat{y_t} = softmax(Wc_t + b) \tag{6}$$

The parameters $W \in \mathbb{R}^{2\times m}$ and $b \in \mathbb{R}^2$ are learned. In order to calculate the loss, we use the cross-entropy as follows:

$$\begin{aligned} &\mathcal{L}(x_1, x_2, .., x_N) = \\ &-\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T^{(n)}}\sum_{t=1}^{T^{(n)}}\Big(y_t^\top log(\hat{y_t}) + (1-y_t)^\top log(1-\hat{y_t})\Big) \end{aligned} \tag{7}$$

The attention mechanism that we used in our model uses MLP[2] to embed each time step's data, and then uses RNN to produce attention weights as in [16] with the difference that the RNN we use is bidirectional, unlike the RETAIN model. On the other hand, the base model of our approach for attention mechanism [22] first encodes words by RNN and then creates attention weights by MLP. In other words, [22] uses the reverse order of RNN and MLP compared to our model. By doing so we maintain the interpretability, using the MLP embedding layer, and then we calculate attentions using RNN, which captures chronological correlations of data and simulates a clinical expert that pays more attention to specific parameters of the data.

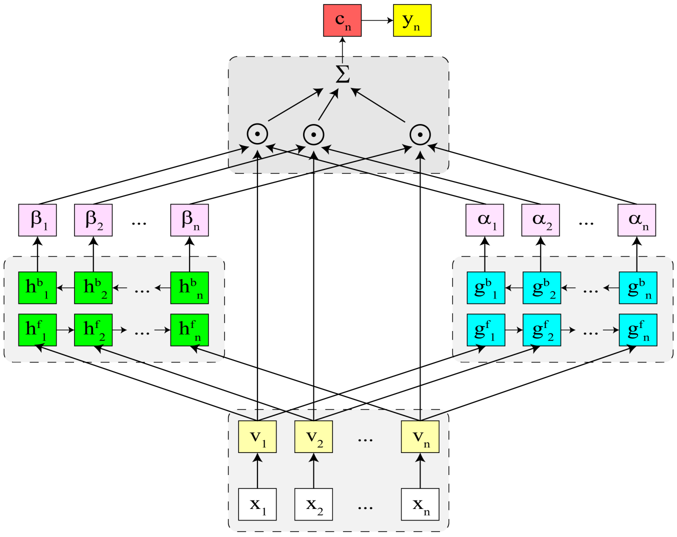

Fig. 1. A conceptual overview of the presented model. The flow of the data through the model can be viewed as a five step procedure. **1)** The inputs are embedded. **2)** The hidden values are calculated. **3)** The attentions are calculated. **4)** The embedded inputs weighted by attentions are summed up to generate contexts. Finally, prediction is made.

C. Interpretation

In order to identify the extent that parameters contribute to a certain result of the prediction, we follow the same approach as RETAIN [16]. α, β and v determine the contributions of each of the medical measurements in our prediction. The idea is that we keep the αs and βs fixed, similar to a physician paying more attention to certain factors she might see more important. Thus, an element in vector x_t which results in highest $y_{n,s}$ has the highest contribution. The predictions are made from the inputs as follows:

$$p(y_n|x_1, ..., x_n) = p(y_n|c_n) = softmax\Big(Wc_t + b\Big) \tag{8}$$

As in equation (5), $c_n \in \mathbb{R}^m$ represents the context vector. The equation (8) can be written as:

$$p(y_n|x_1, ..., x_n) = p(y_n|c_n) = softmax\bigg(W\Big(\sum_{t=1}^{n}\alpha_t\beta_t \odot v_t\Big) + b\bigg) \tag{9}$$

According to equation (1), we can split the vector v_t into $W_{emb} \times x_t$ components. Thus we will have:

$$\begin{aligned} &p(y_n|x_1, ..., x_i) = \\ &= softmax\bigg(W\Big(\sum_{t=1}^{n}\alpha_t\beta_t \odot \sum_{k=1}^{r} x_{t,k}W_{emb}[:, k]\Big) + b\bigg) \\ &= softmax\bigg(\sum_{t=1}^{n}\sum_{k=1}^{r} x_{t,k}\alpha_t W\Big(\beta_t \odot W_{emb}[:, k]\Big) + b\bigg) \end{aligned} \tag{10}$$

$x_{t,k}$ denotes the k-th element of vector x_t. We can rewrite the equation (10) as follows which makes us able to calculate the

[2] Multilayer Perceptron

TABLE I
THE SUMMARY OF PERFORMANCE OF DIFFERENT MODELS ON OUT OF SAMPLE DATA

	AUC			Accuracy		
HL # *	1	2	3	1	2	3
ISAP	0.7478	0.7547	0.7558	0.7484	0.7548	0.7550
RETAIN	0.7539	0.7531	0.7494	0.7540	0.7530	0.7494
CDP-TT	0.6331	0.6265	0.6038	0.6346	0.6285	0.6077

* Hidden Layers Count

contribution of each elements of vector x_t (where $t \leq n$) to the prediction of y_n:

$$\lambda(y_n, x_{t,k}) = \underbrace{\alpha_t W\Big(\beta_t \odot W_{emb}[:,k]\Big)}_{\text{Coefficient of contribution}} x_{t,k} \qquad (11)$$

To make the equation more readable, we omitted the index n from α_t and β_t. However, the attention values α_t and β_t are considered for the step t since we are making a prediction for that step. The coefficient solely represents the amount of contribution since the input vector x_t is binary. We should note that we have discretized the continuous medical measurements x_t. If we were to use the continuous inputs directly with no discretization, the value of λ would be representative of the contribution [16].

V. RESULTS & DISCUSSION

The presented model was trained with multivariate time series of 11700 patients in a 5-fold cross validation. We compared our model with the original RETAIN [16] with different settings and another attention-based approach [18] as baselines. In this section we discuss our results.

A. Baselines and Performances

The baselines for our model are as follows:

- RETAIN model which uses one-directional RNNs with reverse inputs [16]. In order to refer to this model we simply use RETAIN.
- The attention-based approach to capturing disease progression through time proposed in [18]. For referring to this model we use CDP-TT.
- Our model which consist of bi-directional RNNs. And we refer to it by Interpretable Attention-based Sepsis Prediction Model (IASP).

For the RNNs implemented in the above models we consider using 1, 2 and 3 hidden layers. 5734 out of 11700 patients in our cohort are septic. Although it is approximately half of the patients, we consider using ROC curve as the metric of performance [29]. Here there are two classes: septic and non-septic patients. Based on our experiments, our model's performance reaches to over 0.75 accuracy in AUROC. The area under the ROC curve for each of our experiments are listed in table (I). Furthermore, the ROC curves for the different settings of our experiment are depicted in Fig. 3.

Despite the fact that we have implemented the same RETAIN model, presented in [16], in order to compare its performance against our approach, we produced different performance results compared to what the authors have reported. The key difference that explains this contradiction is that we predict the occurrence of a certain disease (in our case, sepsis) at the end of the sequence as in Learning to Diagnose [11]. And in order to feed the continuous medical measurements to the model, we imputed the missing values and then discretized them by changing them to binary vectors. Whereas the data set used in RETAIN and in [18] is in fact of type ICD9 labels and predictions are made for the next time-step of the sequence given the previous encounter sequences.

B. What Do The Graphs Say?

Once the training of the model is done, the contributions for any patient, outside of the training set, can be calculated. For a septic patient, contributions were calculated, normalized, and then visualized as the heat map depicted in Fig. 2. Essentially the heat map illustrated here is composed of several tiles. Each of these tiles represents the amount of the contribution of the medical parameter from its corresponding hour to the final prediction. We should note that the input data is aggregated in hourly bins. The brighter a heat map's tile, the more it contributes to the final prediction of sepsis. The graph's timeline starts from the admission of the patient to the ICU and ends when the patient's data fully complies with the sepsis 3 definition [1].

Our AI model recommends medical doctors to pay more attention to certain parameters that might be more effective on the patient's outcome. Thereupon, according to Fig. 2, we have visualized the real values of parameters with highest effects in Fig. 4. The advantage of our model to the baselines is that since the heat map, illustrated in Fig. 2, is based on the physiological measurements, it provides more details about the correlation of each parameter at its time step and the final predicted outcome. Whereas, this amount of details is not provided by the methods that only use ICD9 codes.

We should note that when training the model, we use both interpolation and forward filling for handling missing values. However, in order to test our model, or for a real world prediction setting, we do not know what will be the next value of a certain variable. Therefore, in order to create these graphs, we only use forward filling for handling missing values. This explains flat segments of the curves in Fig. 4.

VI. CONCLUSION

As deep models advance through time, the *blackbox-ness* of them becomes a major concern. Particularly in health care applications, the need to justify and explain is considerable since physicians might want to know more details about the certain causes of the resulting predictions. So that they could make more proper decisions.

In this research project, we used an attention-based RNN to predict sepsis from the multivariate time series of physiological and clinical labs and measurements. We also calculated the relative effects of each medical parameter to the final outcome of the prediction. However, there are a few shortcomings with

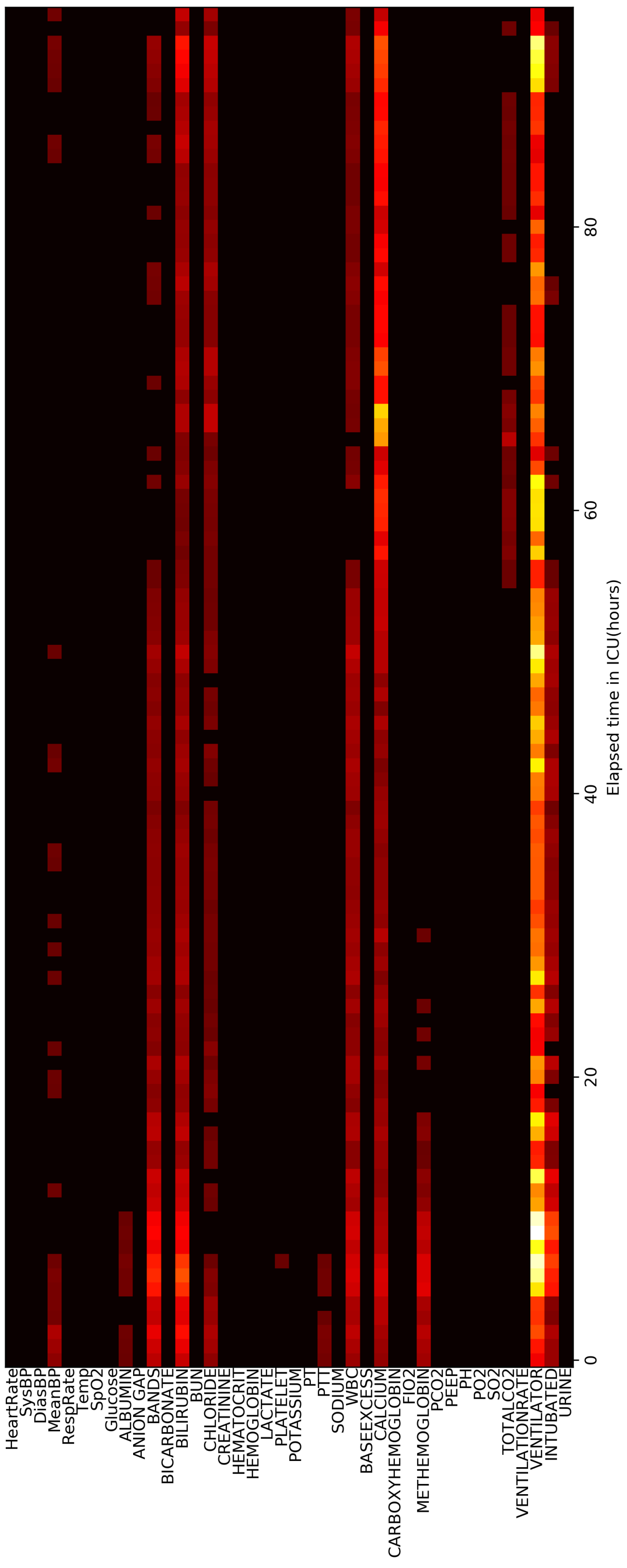

Fig. 2. A heatmap that represents the contributions of each of the measurements along time for a single septic patient. The brighter the tile, the more it contributes to the final prediction that the person is septic. The heatmap starts from the time patient was admitted to ICU up to the point that she is considered to be septic according to sepsis 3 definition.

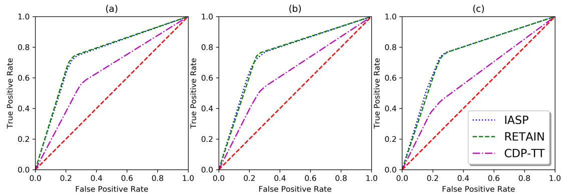

Fig. 3. The ROC curves for different settings of our experiment. Figures (a), (b), and (c) represent the models with 1, 2, and 3 hidden layers in their RNNs,respectively.

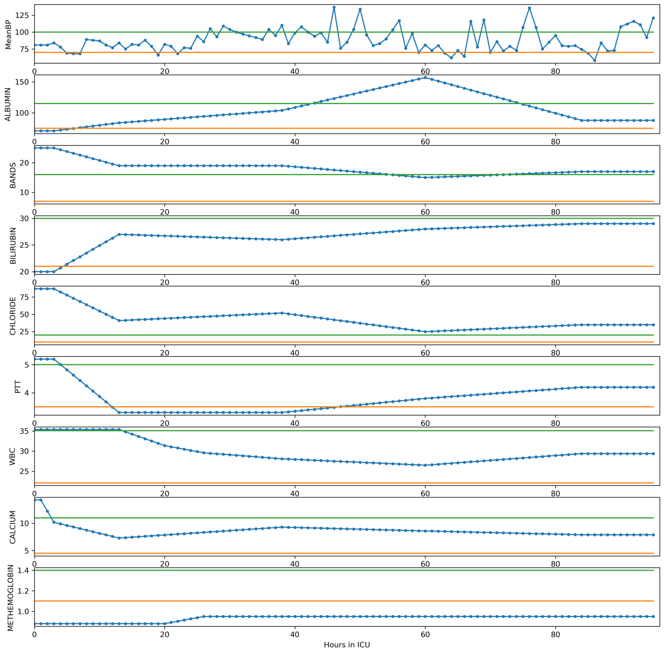

Fig. 4. The raw measurements of parameters with highest contributions for the patient whose contribution graph is depicted in Fig. 2.

our approach which we seek to address in our future work. First of all, it is important to consider evaluating other methods to handle the missing data. Second, a proper clustering of data would possibly yield more interpretability and much robust results. Third, we will explore other ways of feeding the data to our model in order to increase the expressiveness of the dataset which would hopfully yield better performance metrics.

Acknowledgment

The authors would like to thank Dr. Tian Bai from Temple University, Dr. Alistair Johnson from Massachusetts Institute of Technology, and Dr. Sean Bozorgzad, MD from PeaceHealth United General Medical Center, who provided insight and expertise that assisted this research project.

References

[1] S. M, D. CS, S. C, and et al, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[2] A. E. W. Johnson, J. Aboab, J. D. Raffa, T. J. Pollard, R. Deliberato, L. A. Celi, and D. Stone, "A comparative analysis of sepsis identification methods in an electronic database," *Critical Care Medicine*, vol. 46, p. 1, 01 2018.

[3] F. Khoshnevisan, J. S. Ivy, M. Capan, R. Arnold, J. Huddleston, and M. Chi, "Recent temporal pattern mining for septic shock early prediction," in *IEEE International Conference on Healthcare Informatics, ICHI 2018, New York City, NY, USA, June 4-7, 2018*, pp. 229–240, IEEE Computer Society, 2018.

[4] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, pp. 707–710, Jul 1996.

[5] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "Apache ii: a severity of disease classification system.," *Critical care medicine*, vol. 13 10, pp. 818–29, 1985.

[6] C. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *QJM: An International Journal of Medicine*, vol. 94, no. 10, pp. 521–526, 2001.

[7] F. van Wyk, A. Khojandi, R. Kamaleswaran, O. Akbilgic, S. Nemati, and R. L. Davis, "How much data should we collect? a case study in sepsis detection using deep learning," in *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pp. 109–112, Nov 2017.

[8] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *CoRR*, vol. abs/1606.01865, 2016.

[9] J. Futoma, S. Hariharan, and K. A. Heller, "Learning to detect sepsis with a multitask gaussian process RNN classifier," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1174–1182, PMLR, 2017.

[10] J. Futoma, S. Hariharan, K. A. Heller, M. Sendak, N. Brajer, M. Clement, A. Bedoya, and C. O'Brien, "An improved multi-output gaussian process rnn with real-time validation for early sepsis detection," in *MLHC*, 2017.

[11] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *CoRR*, vol. abs/1511.03677, 2015.

[12] S. P. Shashikumar, M. D. Stanley, I. Sadiq, Q. Li, A. L. Holder, G. D. Clifford, and S. Nemati, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics.," *Journal of electrocardiology*, vol. 50 6, pp. 739–743, 2017.

[13] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," *Comput. Biol. Med.*, vol. 74, pp. 69–73, July 2016.

[14] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016.

[15] E. Choi, M. T. Bahadori, and J. Sun, "Doctor AI: predicting clinical events via recurrent neural networks," *CoRR*, vol. abs/1511.05942, 2015.

[16] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism," *CoRR*, vol. abs/1608.05745, 2016.

[17] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, (New York, NY, USA), pp. 233–240, ACM, 2017.

[18] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining*, KDD '18, (New York, NY, USA), pp. 43–51, ACM, 2018.

[19] B. C. Kwon, M. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *CoRR*, vol. abs/1805.10724, 2018.

[20] X. Wang, Z. Wang, J. Weng, C. Wen, H. Chen, and X. Wang, "A new effective machine learning framework for sepsis diagnosis," *IEEE Access*, vol. 6, pp. 48300–48310, 2018.

[21] S. Nemati, A. L. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the icu.," *Critical care medicine*, vol. 46 4, pp. 547–553, 2018.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[23] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[24] K. J Carroll, "On the use and utility of the weibull model in the analysis of survival data," *Controlled clinical trials*, vol. 24, pp. 682–701, 01 2004.

[25] J. Alistair E.W., P. Tom J., S. Lu, L. Li-wei H., F. Mengling, G. Mohammad, M. Benjamin, S. Peter, A. C. Leo, and M. Roger G., "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, 2016.

[26] A. Johnson and T. Pollard, "sepsis3-mimic," May 2018.

[27] R. KE, S. CW, A. AR, and et al, "Association of the quick sequential (sepsis-related) organ failure assessment (qsofa) score with excess hospital mortality in adults with suspected infection in low- and middle-income countries," *JAMA*, vol. 319, no. 21, pp. 2202–2211, 2018.

[28] A. Kratz, M. Ferraro, P. M. Sluss, and K. B. Lewandrowski, "Normal reference laboratory values," *New England Journal of Medicine*, vol. 351, no. 15, pp. 1548–1563, 2004. PMID: 15470219.

[29] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.