Dynamic Illness Severity Prediction via Multi-task RNNs for Intensive Care Unit

Weitong Chen
The University of Queensland
Brisbane, Australia
w.chen9@uq.edu.au

Sen Wang
Griffith University
Gold Coast, Australia
sen.wang@griffith.edu.au

Guodong Long
University of Technology Sydney
Sydney, Australia
guodong.long@uts.edu.au

Lina Yao
The University of New South Wales
Sydney, Australia
lina.yao@unsw.edu.au

Quan Z. Sheng
Macquarie University
Sydney, Australia
michael.sheng@mq.edu.au

Xue Li
The University of Queensland
Nanjing University of Aeronautics and Astronautics
xueli@itee.uq.edu.au

***Abstract*—Most of the existing analytics on ICU data mainly focus on mortality risk prediction and phenotyping analysis. However, they have limitations in providing sufficient evidence for decision making in a dynamically changing clinical environment. In this paper, we propose a novel approach that simultaneously analyses different organ systems to predict the illness severity of patients in an ICU, which can intuitively reflect the condition of the patients in a timely fashion. Specifically, we develop a novel deep learning model, namely MTRNN-ATT, which is based on multi-task recurrent neural networks. The physiological features of each organ system in time-series representations are learned by a single long short-term memory unit as a specific task. To utilize the relationships between organ systems, we use a shared LSTM unit to exploit the correlations between different tasks for further performance improvement. Also, we apply an attention mechanism in our deep model to learn the selective features at each stage to achieve better prediction results. We conduct extensive experiments on a real-world clinical dataset (MIMIC-III) to compare our method with many state-of-the-art methods. The experiment results demonstrate that the proposed approach performs better on the prediction tasks of illness severity scores.**

***Index Terms*—deep learning, multi-task learning, clinical informatics, illness severity prediction.**

I. Introduction

The accumulation of more than 18.8 million electronic health record (EHR) available in the *My Health Record System*[1] has attracted a great attention research attention from data scientists. In the intensive care unit (ICU),), there is an abundance of multi-format electronic data on patients in terms of data types and recording frequency, but this data is vendor-specific and limited in scope [1]. Learning such a large volume of data from different sources could provide strong supporting evident for decision making in an ICU which could have a beneficial effect on clinical practice. Clinical decision making in the ICU is fundamentally driven by predicting an outcome for the patient of sustained benefit in terms of quality and length of life.

The recent advances in deep learning methods have enable computer scientists to develop innovative ICU-based decision support systems. However, they mainly concentrate on mortality prediction [2], [3] and phenotyping analysis [4]–[6]. and fail to provide clinicians with important physiological insight to support decision making in a dynamically changing clinical environment. In an ICU, numerous scoring systems, such as SOFA Score [7], APACHE II Score [3], SAPS II Score [8], etc., can reflect the illness severity of patients with sufficient medical insights on different aspects. Unfortunately, these evaluations are conducted in a long time window, such as 24 hours, resulting in low level of responsiveness to critically ill patients. Inspired by early warning scoring systems [9], we believe that a more frequent routinely evaluation of a critical ill patient can clearly visualise the patient's' condition providing a potential pathway to achieve the early prediction of a medical condition. In Fig 1, two SOFA trajectories of patients in the ICU are depicted, showing that an hourly SOFA score can be used as the immediate ground truth of a patient's conditions in the ICU. Hence, predicting the illness severity score in a denser manner is an effective solution

This work is partially supported by ARC project ID: DP160104075, Project ID: LP150100671 partnership with ARACY and GBCA, project ID: LP160100630 partnership with Australia Government Department of Health.

[1] https://myhealthrecord.gov.au

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00111

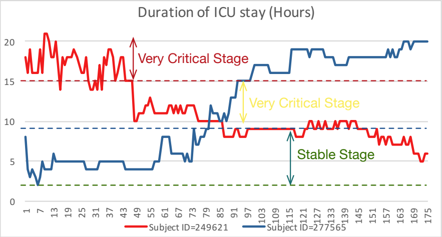


Fig. 1: The figure shows two SOFA trajectories of patients in an ICU. The patient (indicated by the red line) was in a critical condition when initially admitted to the ICU, but progressively improves and is eventually discharged from the ICU. In contrast, the condition of the patient (measured by the SOFA score and indicated by the blue line) deteriorates during his initial ICU stay, and eventually dies in hospital. The X-axis is the duration of stay (in hours) and the Y-axis is the SOFA score.

for the prompt monitoring of patients in the ICU. Over the past few years, AI researchers have made efforts to develop various deep models for health data. As ICU data are mainly represented by multivariate time-series, recurrent neural networks (RNNs) have been investigated to achieve better performance in mortality prediction and disease code prediction. ICU data are often sparse, irregularly sampled, and noisy. Many imputation methods have been studied and incorporated into the learning framework to achieve better performance.

On the other hand, it is commonly known that the organ systems in the human body are biologically correlated in clinical trials. This has been largely ignored by AI researchers as the majority of existing deep models treat all physiological time-series features as the entire input from the human body without considering the correlation between separate organ systems, which may be detrimental to prediction performance. This may be because the existing work is not sensitive to only a few physiological features that are abnormal when the related organ starts to fail.

To address these problems, in this paper, we propose a novel approach that simultaneously analyses different organ systems to predict the illness severity of patients in an ICU, intuitively displaying the condition of the patients in a timely manner. The proposed approach makes the following contributions:

- We propose a novel multi-task RNN framework to simultaneously learn each human organ system in consideration of the temporal correlations between organ systems by a shared unit. In this way, the performance is further improved. To the best of our knowl- edge, this work is the first to analyse ICU patient data systematically.
- To learn the selective features, we apply the attention mechanism to each learning unit to focus on important information in the input to achieve better performance.
- Extensive experiments are conducted by comparing many state-of-the-art methods with our pro- posed approach on a real-world ICU dataset respect to illness severity prediction. The results show that our method outperforms the compared methods in the prediction of illness severity.

The remainder of this paper is organized as follows: the related work is reviewed in Section II. In Section III, we describe our proposed method in detail, followed by the experiment evaluations on a real-world dataset and related discussions in Section IV. We conclude this paper in Section V.

II. Related Work

In intensive care units, the EHRs of patients mainly consist of time-series data storing various types of medical variables. Thus, time-series models have been prioritized in EHR data analysis for different learning tasks, such as patient mortality prediction, disease diagnosis, and disease progression modelling.

To address the these analytic problems using EHRs, over the past few years, researchers have developed many sophisticated models to analyze medical data over the past few years. Ghassemig et al. [10] and Wang et al. [4] exploit the correlations embedded in features to predict the development and phenotype of ICU patients. To improve the prediction performance, Nie et al. [11] and Zhou et al. [12] take both the consistency in the prediction results among multiple modalities and the selected task-specific features into consideration to solve prediction problems. Unfortunately, the performance of these methods is limited due to the underutilization of temporal information within the time-series. Temporal features that reflect changes in the patients condition in a timely manner can improve the performance on such prediction tasks. Using a monotonically decreasing function, Pham et al. [13] modified the standard LSTM model to handle irregular time-series data to predict medicine for patients. Similarly, Che et al. [14] and Lipton et al. [6] add time interval terms based on GRU and LSTMs to learn irregular EHR time-series data in health-care applications. However, these works process the “time” information in a heuristic manner, by using a monotonically decreasing function. Hence, these methods may cause over- or under-parameterisation in modelling time intervals [15].

Multi-task Learning (MTL) aims to leverage useful information contained in multiple related tasks to improve the overall performance. It has been widely studied in computer vision [16]–[18]. An et. al employed a multi-task convolutional neural network (CNN) for image-based multi-label attribute prediction, in which CNNs are used for capturing features from image segments and a common layer is responded to determining the inter-relationships of images [19]. On the other hand, Li et. al [20] proposed a heterogeneous multi-task learning framework to exploit the correlations between the various visual learning tasks by jointly learning the image features for two different prediction tasks. The key idea of multi-task learning in clinical analytics is to capture the intrinsic relationships between medical tasks. By utilizing the common set of features significant to all tasks, Zhou et al. [21], target at predicting disease progression. In addition, Harutyunyan [22] extends the success of heterogeneous model for clinical time-series learning. However, the performance of this model result in over-parametrisation in modelling.

The Attention mechanism has been successfully applied in many learning tasks related to natural language processing [23]–[25]. Bahdanau et al. [26], extended the basic encoder-decoder architecture with the attention mechanism to allow the model to select parts of the feature vectors that are relevant to predicting a target to achieve competitive performance. In [27], Vinyals et al. compute an attention with weights, which reflect how much attention should be put over the input vectors, boosting the performance on a large scale. Sharma et al. [28] applied a location-based softmax function to the hidden states of the LSTM layers, thus recognizing more valuable elements in sequential inputs for action recognition.

III. Proposed Method

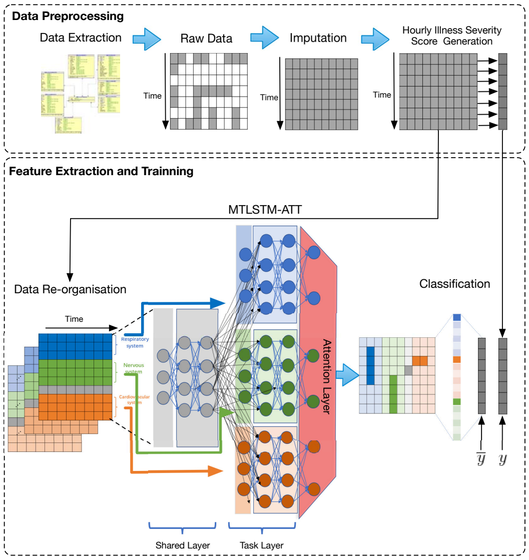


Fig. 2: The work flow of the proposed approach

In this section, we present the details of our novel framework, as illustrated in Fig. 2 , based on multi-task deep RNNs with attention mechanism to predict the illness servility of ICU patient using SOFA scores. Firstly, we introduce data prepossessing including the cohort selection, data extraction, data cleaning and feature extraction methods. We then describe the detailed architecture of the proposed multi-task model, which is not only able to learn the distinctive features from different human organ systems in time-series EHRs, but can also exploit the temporal correlation between organ systems. Lastly, we explain how we embed the attention mechanism, which focuses on selectively recognizing features, into the proposed multi-task framework to learn descriptive representations for better perdiction results.

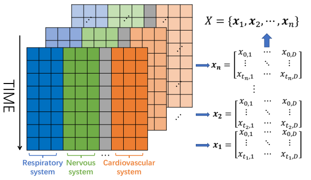

TIME
Respiratory system
Nervous system
Cardiovascular system
$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$
$\boldsymbol{x}_n = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_n,1} & \cdots & x_{t_n,D} \end{bmatrix}$
$\boldsymbol{x}_2 = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_2,1} & \cdots & x_{t_2,D} \end{bmatrix}$
$\boldsymbol{x}_1 = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_1,1} & \cdots & x_{t_1,D} \end{bmatrix}$

Fig. 3: Illustration of data structure and organization.

A. Data Reprocessing

We applied the latest version (v1.4) of MIMIC-III [29], which contains 53,423 de-identified adult admissions between 2001 and 2012. Following the convention, we only extract patients who are adults, age between 16 and 75 (16≤age ≤75), and who stayed in the ICU less than 12 hours. For each adult patient, multiple admissions with many ICU stays may exist. In this work, we treat each ICU stay as an independent data observation, which result in 36,740 records. To collect multi-variate physiological features, we have extracted 41 features[2] with respect to different human organ systems from multiple tables in MIMIC III. The values of each feature within a time window are averaged as the new value in that time slot. For each stay record, all the extracted features will be converted into a matrix with a variable number of rows, as illustrated in Fig. 3. D is the number of features, while n is the number of ICU stay records. We use t_i to denote the max length in time for the i-th data sample, $i = 1, \cdots, n$. In this way, the data samples can be represented by $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^{t_i \times D}$.

As pointed out in [30], the extracted data does not of having good quality due to noise, missing values, outliers, etc. We have borrowed the same procedures in [30] to improve the data quality. For missing value of the d-th variable at time t, we have used the *forward-fill* imputation strategy in [6] as follows, 1) if there is at least one valid observation at time $t' < t$, we set $x_{t,d} := x_{t',d}$. 2) if there are no previous recorded values, then we replace the missing value with the median value over all measurements. This strategy is inspired by the fact

[2] https://github.com/AnthonyTsun/List of 41 features III.pdf

that measurements are recorded at intervals proportional to the rate at which the values are believed or observed to change.

B. Multi-Task Recurrent Neural Networks

The recurrent neural networks (RNNs) [31] have sufficient ability to process arbitrary sequential inputs by recursively applying a transaction function to its *hidden vector* $\mathbf{h}_t$. However, RNNs have difficulties learning long-range dependencies. The LSTM network [32] has been proposed to address the vanishing gradient problem by incorporating gating functions to control state updates and outputs [33]. The LSTM at each time step t includes i, f, o, c and h, which are the input gate, forget gate, output gate, memory cell, and hidden state respectively. The forget gate controls the amount of memory to be "forgotten", while the input gate controls the update of each unit and the output gate controls the exposure of the cell state. The LSTM can be defined as follows:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\
c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
h_t &= o_t \tanh(c_t),
\end{aligned} \tag{1}$$

where x_t is the input at a time t, W are weights, bs are the bias terms, and σ denotes the logistic sigmoid function. To learn features with respect to each organ system, we have modified the LSTM for each task and added a shared layer to exploit the temporal correlations between different systems. Fig. 4 shows the structure of our multi-task LSTMs. Specifically, features from different systems are fed into different LSTMs. For example, features denoted by $\boldsymbol{x}$ from the respiratory system and the ones denoted by $\boldsymbol{x'}$ from cardiovascular system are simultaneously fed into two separate LSTMs, i.e., $LSTM^{(m)}$ and $LSTM^{(n)}$, each of which is regarded as a different task and aims to captures the intrinsic features in long-short terms, respectively. As human organs work collaboratively together, it is believed that there must be correlations between organ systems, which can be beneficial to learning tasks. To capture such kinds of temporal correlation between systems, we introduced a shared layer $LSTM^{(s)}$, as shown in the middle of Fig. 4, in our framework. The shared hidden layer fully connects with all the other LSTMs layers, e.g., $LSTM^{(m)}$ and $LSTM^{(n)}$ in the figure. The activation function f of the current hidden state for the shared layer, $\mathbf{h}_t^{(s)}$, is the same as the one in Eq. (1). In contrast, we have modified the activation function for each LSTM $\boldsymbol{h}_t^{(m)}$, which learns different organ features as follows:

$$\mathbf{h}_t^{(m)} = \begin{cases} 0 & t = 0 \\ f\left(\mathbf{h}_{t-1}^{(m)} \odot \mathbf{h}_{t-1}^{(s)}, x_t^{m,i}\right) & otherwise, \end{cases} \tag{2}$$

where $\odot$ denotes a concatenate operation. Meantime, We also change the state $c_t^{(m)}$ for in each task-specific LSTM ($LSTM^{(m)}$ or $LSTM^{(n)}$) as follows:

$$\begin{aligned}
c_t^{(m)} = f_t c_{t-1}^{(m)} + i_t^{(m)} \tanh(W_{xc}x_t^{m,i} \\
+ W_{hc}h_{t-1}^{(m)} \odot h_{t-1}^{(s)} + b_c^{(m)}),
\end{aligned} \tag{3}$$

where $x_t^{m,i}$ is the input at time t. $h_{t-1}^{(m)}$ is the output of Eq. (2) when $t-1$. The shared hidden layer outputs $h_{t-1}^{(s)}$ when $t-1$.

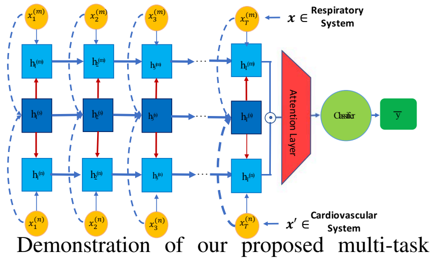


Fig. 4: Demonstration of our proposed multi-task LSTM architecture with a shared hidden layer.

C. Attention Mechanism

Attention mechanisms have been widely adopted in computer vision and natural language to "focus on" a certain region of information while perceiving the surrounding information, and then adjusting the focal point over time. Taking the advantages of the attention mechanism, we incorporate the attention mechanism into our model to selectively learn the important features. As shown in Fig. 4, we have added one attention function into our proposed multi-task model, For each patient, we calculate the attention weight using the dot-product of the hidden state for every feature in the input. Thus the score for the t-th feature $score_t$ is calculated as follows:

$$score_t = h_f^\top \hat{h}_s \tag{4}$$

where $\hat{h}_s$ is the concatenated hidden state of RNN, in which the t-th feature of the input is imputed, and h_f is the learned feature of the input. By using the score, the weight for the t-th feature W_t can be computed as follows:

$$W_t = \frac{\exp\left(score_t\right)}{\sum_{t'} \exp\left(score_{t'}\right)} \tag{5}$$

where, t' denotes all the features in the input. By using the weight and the hidden state of the feature, the final state output a_f is computed as a convex sum of hidden states h_t:

$$a_f = \sum W_t h_t \tag{6}$$

The reasons for adopting the attention mechanisms are two-fold: 1) the attention function gives high weight to the feature which strongly affects the vector representation of the whole input; and 2) it establishes direct short-cut connections between the target and the source.

IV. Experiments

A. Dataset Description and Experiment Design

To evaluate the performance of the proposed model, we have conducted extensive experiments on a publicly available benchmark dataset *MIMIC III*[3] and compared the proposed approach with several baselines and many state-of-the-art algorithms.

[3] https://mimic.physionet.org/

TABLE I: Demonstration of SOFA score and corresponding classes.

Sofa Score	0-6	7-9	10-12	13-14	15	15-24
Mortality Rate	10%	15 - 20%	40-50%	50-60%	80%	90%
Class	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6

We have tested all the methods on the latest version (v1.4) of the MIMIC-III dataset [29]. We only focused on adult patients who are older than and equal to 16 years old but younger than and equal to 75 years old ($16 \leq \text{age} \leq 75$). Also, we excluded those records where the length of the ICU stay is less than 12 hours. To train a supervised model, we have used an entropy-based loss function with a ℓ_2-norm regularization term. We randomly selected 80% of 36,740 ICU stays as training data, while the remaining 20% of the data were used as testing samples. To select best parameters, we have employed a 5-fold cross-validation schema in the experiments. All the experiments are repeatedly run 10 times. Our implementation is available at GitHub [4].

B. Comparison Methods

We evaluate the effectiveness of MTRNN-ATT by comparing it with the following state-of-the-art approaches and baselines regarding AUC, precision, recall, and F1-Score.

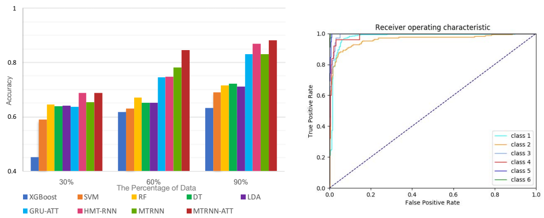
1) **GRU-ATT**: Nguyen et al. [34] have employed a Gated Recurrent Unit (GRU) and the attention mechanism for illness servility prediction.
2) **HMT-RNN**: Hharutyunyan et al. [22] have developed a framework based on RNNs to predict in-hospital mortality.
3) **pRNN**: Aczonet al. [35] have investigated pediatric encounter records(physiologic observations, laboratory results, drugs, and interventions) with a RNN framework for mortality prediction.
4) **RNN**: We have implemented a standard version of single task recurrent neural network with hyperparameter suggested by Hharutyunyan et al. [34] as one of baselines.
5) **MTRNN**: We have implemented a multi-task recurrent neural network without the attention mechanism [36].
6) **RNN-ATT**: We constructed a signal task recurrent neural network [34] with the attention mechanism as another baseline method.

In addition to the the set of state-of-the-arts, we also compare our proposed method with other baseline classification methods, namely **Support Vector Machine** (SVM), **Random Forest** (RF), **Decision Tree** (DT), **Linear discriminant analysis** (LDA), and **XGboost** [5]. For comparisons purpose, we keep all the hyper-parameter of the baseline models the same.

The SOFA scoring system is useful in predicting the clinical outcomes of critically ill patients. The estimation of mortality risk is based on the maximum (highest) SOFA score during a patient's ICU stay as shown in Table I [7]. We group the SOFA scores into six classes as shown in the TableI. To consider the classification performance, we report the results of all the methods measured by AUC in Table II. It is clear that our approach outperforms all the compared methods. For example, our method has an improvement of 3% in accuracy over the second best method (#2) [22]. We can observe that multi-task RNNs (#11) performs better than most of their counterparts with single task settings. Although the baseline MT-RNN (#10) focused on multi-task learning, our model surpasses their method significantly. It may contribute to the exploitation of the temporal correlations between different human organ systems by the shared hidden layer. Furthermore, the attention mechanism can boost the performance.

TABLE II: Comparison between our method and all the compared methods.

Index	Method	Accuracy
1	GRU-ATT	0.8305
2	HMT-RNN	0.8690
3	pRNN	0.8041
4	SVM	0.6893
5	RF	0.7153
6	DT	0.7230
7	LDA	0.7122
8	XGBoost	0.6334
9	RNN	0.8010
10	MT-RNN	0.8630
11	RNN-ATT	0.8303
12	**MTRNN-ATT**	**0.8990**



(a) Performance of prediction on sub-sampled experiment dataset. X-axis, sub-sampled dataset size; Y-axis, accuracy.

(b) The ROC curves showing the discrimination capability of a classifier. X-axis, false positive rate; Y-axis, true positive rate.

TABLE III: Evaluation on Different Time Windows

(a) Precision **(b)** Recall

Index	Method	1hr	3hr	6hr	Index	Method	1hr	3hr	6hr
1	**SVM**	0.6929	0.6317	0.5412	1	**SVM**	0.7240	0.6360	0.5741
2	**RF**	0.7059	0.6719	0.6251	2	**RF**	0.7373	0.6834	0.6334
3	**DT**	0.7302	0.6526	0.6095	3	**DT**	0.7341	0.6731	0.6294
4	**LDA**	0.7021	0.6528	0.6010	4	**LDA**	0.7183	0.6590	0.6517
5	**XGBoost**	0.6410	0.6173	0.4525	5	**XGBoost**	0.6422	0.6288	0.5192
6	**RNN**	0.7899	0.8063	0.6362	6	**RNN**	0.8182	0.8261	0.6354
7	**MT-RNN**	0.8671	0.8124	0.6903	7	**MTRNN**	0.8713	0.8613	0.8015
8	**RNN-ATT**	0.8303	0.8062	0.7215	8	**RNN-ATT**	0.8411	0.8223	0.7021
9	**MTRNN-ATT**	**0.8886**	**0.8783**	**0.8202**	8	**MTRNN-ATT**	**0.8972**	**0.8623**	**0.8320**

In Fig. 5b, we have depicts the ROC curves of our methods. From the figure, we can observe that our method is very sensitive to predictions of Class 1 and Class 6, which are two extreme cases in the ICU. This means that the proposed method can effectively forecast the very critical conditions on the way for the patients. However, our method can also well discriminate the intermediate conditions. We present the experiment results with respect to different time window lengths of 1-hour, 3-hours, and 6-hours. Table IIIa and Table IIIb have shown the better performance of our method over

[4] https://github.com/AnthonyTsun/MultiTask_Attention

[5] Baseline methods were implemented by using the scikit-learn toolkit.

TABLE IV: Evaluation of the influence of the Attention Mechanism and the Multi-Task Model

	Method	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Average
Percision	RNN	0.8531	0.6562	0.5294	0.5714	0.7058	0.6875	0.7899
	MT-RNN	0.9212	0.7785	0.6470	0.5714	0.7667	0.7567	0.8671
	RNN-ATT	0.8920	0.7021	0.6667	0.3000	0.7575	0.8461	0.8303
	MTRNN-ATT	**0.9391**	**0.7972**	**0.6705**	**0.5769**	**0.7778**	**0.9667**	**0.8886**
Recall	RNN	0.9756	0.3705	0.5124	0.2514	0.6153	0.6470	0.8112
	MT-RNN	0.9569	0.6705	0.6285	0.6400	0.5897	0.8265	0.8713
	RNN-ATT	0.9598	0.5823	0.5714	0.2400	0.6412	0.6475	0.8411
	MTRNN-ATT	**0.9741**	**0.6825**	**0.6857**	**0.7667**	**0.7567**	**0.8529**	**0.8972**
F1	RNN	0.9103	0.4768	0.5217	0.2500	0.6573	0.6667	0.7878
	MT-RNN	0.9387	0.7272	0.6376	0.5882	0.6667	0.7883	0.867
	RNN-ATT	0.9246	0.6366	0.6153	0.2667	0.6944	0.7333	0.9329
	MTRNN-ATT	**0.9563**	**0.7284**	**0.6437**	**0.6772**	**0.7466**	**0.9063**	**0.8878**
AUC	RNN	0.8956	0.8221	0.9279	0.9303	0.9591	0.9634	0.9329
	MT-RNN	0.9370	0.8961	0.9738	0.9716	0.9483	0.9744	0.9497
	RNN-ATT	0.9195	0.8743	0.9683	0.9356	0.9498	0.9551	0.9415
	MTRNN-ATT	**0.9631**	**0.9174**	**0.9807**	**0.9755**	**0.9886**	**0.9829**	**0.9761**

the baseline methods. Also, we have observed that with an increase of time window, the prediction performance slightly drops. This may be because the variations in medical conditions can change in different directions, either worse or better, after a longer period of time. Table IV illustrates the classification performance with respect to each class. From the results in Table IV, it can be seen that our method performs better result in most cases than all the baseline methods with respect to each class in terms of various metrics, including Precision, Recall, and F1-Score.

V. Conclusion

In this paper, we proposed a novel deep learning framework that simultaneously analyses different human organ systems to predict the illness severity of patients in the ICU. Our framework is based on multi-task LSTMs and treats each organ system separately and also exploits the correlations between organ systems by a shared unit. To the best of our knowledge, this work is the first to analyse ICU patients systematically. To deal with problems raised by data quality, we applied the attention mechanism to give a higher weight to the important features of the input to further improve the models performance. Through comprehensive experiments, we showed that our approach outperforms all the compared methods and baselines on illness severity prediction.

References

[1] H. Binder and M. Blettner, "Big data in medical science," *Deutsches ÄI*, 2015.

[2] F. Shann, G. Pearson, A. Slater, and K. Wilkinson, "Paediatric index of mortality (pim): a mortality prediction model for children in intensive care," *Intensive care medicine*, pp. 201–207, 1997.

[3] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano *et al.*, "The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[4] S. Wang, X. Chang, X. Li *et al.*, "Diagnosis code assignment using sparsity-based disease correlation embedding," *TKDE*, 2016.

[5] S. Wang, X. Li, L. Yao, Q. Z. Sheng, G. Long *et al.*, "Learning multiple diagnosis codes for icu patients with local disease correlation mining," *TKDD*, 2017.

[6] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with rnns," *Machine Learning for Healthcare*, 2016.

[7] J. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive care medicine*, 1996.

[8] 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*, pp. 2957–2963, 1993.

[9] D. C. Bouch and J. P. Thompson, "Severity scoring systems in the critically ill," *Critical Care & Pain*, 2008.

[10] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *KDD*, 2014.

[11] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua, "Beyond doctors: Future health prediction from multimedia and multimodal observations," in *ACMM*, 2015.

[12] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *KDD*, 2012.

[13] T. Pham, T. Tran *et al.*, "Deepcare: A deep dynamic memory model for predictive medicine," in *PAKDD*. Springer, 2016.

[14] Z. Che, S. Purushotham *et al.*, "Recurrent neural networks for multivariate time series with missing values," *arXiv*, 2016.

[15] K. Zheng, J. Gao, K. Y. Ngiam, B. C. Ooi, and W. L. J. Yip, "Resolving the bias in electronic medical records," in *KDD*, 2017.

[16] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *CVPR*, 2015.

[17] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *IJCV*, 2013.

[18] Z. Hong, X. Mei *et al.*, "Tracking via robust multi-task multi-view joint sparse representation," in *CICCV*. IEEE, 2013.

[19] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *TMM*.

[20] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," in *CVPR*, 2014, pp. 482–489.

[21] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *KDD*, 2011.

[22] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[23] T. Rocktäschel, E. Grefenstette *et al.*, "Reasoning about entailment with neural attention," *arXiv*, 2015.

[24] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *HLT*, 2016, pp. 1480–1489.

[25] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *NLP*, 2015, pp. 1422–1432.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv*, 2014.

[27] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *NIPS*, 2015.

[28] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv*, 2015.

[29] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, 2016.

[30] S. Purushotham, C. Meng *et al.*, "Benchmark of deep learning models on large healthcare mimic datasets," *arXiv*, 2017.

[31] J. L. Elman, "Finding structure in time," *Cognitive science*, 1990.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, pp. 1735–1780, 1997.

[33] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[34] P. Nguyen, T. Tran, and S. Venkatesh, "Deep learning to attend to risk in icu," *arXiv*, 2017.

[35] M. Aczon, D. Ledbetter, L. Ho, A. Gunny, A. Flynn, J. Williams, and R. Wetzel, "Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks," *arXiv*, 2017.

[36] W. Chen, S. Wang *et al.*, "Eeg-based motion intention recognition via multi-task rnns," in *SIAM DM*, 2018.