



AMRNN: attended multi-task recurrent neural networks for dynamic illness severity prediction

Weitong Chen¹ · Guodong Long¹ · Lina Yao² · Quan Z. Sheng³

Received: 17 March 2019 / Revised: 23 July 2019 / Accepted: 6 August 2019 /

Published online: 21 September 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Illness severity prediction (ISP) is crucial for caregivers in the intensive care unit (ICU) while saving the life of patients. Existing ISP methods fail to provide sufficient evidence for the time-critical decision making in the dynamic changing environment. Moreover, the correlated temporal features in multivariate time-series are rarely be considered in existing machine learning-based ISP models. Therefore, in this paper, we propose a novel interpretable analysis framework which simultaneously analyses organ systems differentiated based on the pathological and physiological evidence to predict illness severity of patients in ICU. It not only timely but also intuitively reflects the critical conditions of patients for caregivers. In particular, we develop a deep interpretable learning model, namely AMRNN, which is based on the Multi-task RNNs and Attention Mechanism. Physiological features of each organ system in multivariate time series are learned by a single Long-Short Term Memory unit as a dedicated task. To utilize the functional and temporal relationships among organ systems, we use a shared LSTM task to exploit correlations between different learning tasks for further performance improvement. Real-world clinical datasets (MIMIC-III) are used for conducting extensive experiments, and our method is compared with the existing state-of-the-art methods. The experimental results demonstrated that our proposed approach outperforms those methods and suggests a promising way of evidence-based decision support.

Keywords Multi-task learning · Deep learning · Illness severity prediction

This article belongs to the Topical Collection: *Special Issue on Application-Driven Knowledge Acquisition*

Guest Editors: Xue Li, Sen Wang, and Bohan Li

✉ Weitong Chen
w.chen9@uq.edu.au

Extended author information available on the last page of the article.

1 Introduction

The accumulation of over 69 million EHRs from over 21.9 million (90.1% of National Participation Rate) individuals in the *My Health Record System*¹, has caught great attentions from machine learning and data mining communities. Learning such a large volume of data from different sources could provide strong supports for evidence-based clinical decision making in ICU, which could benefit clinical practice. However, multi-format of EHRs are abundant in terms of data categories, data types, and multivariate time series, but are usually vendor-specific and limited in scope [3]. A clinical decision in ICU is fundamentally driven by forecasting an outcome for patients in terms of quality and length of life [3]. Recently, deep learning technics have advanced the researches in ICU decision support. But, they mainly focus on mortality estimation [23] and phenotype analysis [16]. In general, clinical decisions in ICUs are time-critical and highly dependant on the physiological data analysis. Without sufficient real-time patients' information, making an accurate and rapid decision would be very challenging for clinicians in a fast-changing environment. In the past decades, numerous scoring systems, such as SOFA [26], APACHE II [14], SAPS II [15], etc., have been introduced and progressively refined to assess conditions of patients and evaluate the effectiveness of treatment in ICU. These scores can reflect the status of patients with adequate medical insights from different physiological aspects.

1.1 Motivation

Although numerous scoring systems can reflect the critical condition of patients on different physiological aspects, they are unable to provide time-critical information because they are currently still a hand-crafted calculation with daily time intervals. Since most of the medication evaluation, such as pathology tests, are organized over a long-period of time-window, e.g. 24-hours, it results in less responsiveness to patients who are in critical condition. Referring to the need for providing early warning scoring systems [4] for ICU patients, we believe that an instant scoring system can help monitor the medical development of patients. To demonstrate this idea, we firstly conducted a study on showing how medical conditions are developed with two SOFA score trajectories for two patients. As it can be seen in Figure 1, two SOFA trajectories of patients can envision the development of diseases and the effectiveness of treatments. The high-frequency SOFA scores can be used as a baseline of patients' medical conditions. In this way, continuously predicting the illness severity scores can be a new tool for monitoring patients.

1.2 Challenge

ICU data is mainly recorded as multivariate time series. Over the years, RNNs (Recurrent Neural Networks) and their variants that are used as a deep model capable of capturing features of time series, have been investigated and achieved significant results on the mortality prediction and disease-code prediction. In this field of research, ICU data is characterized as sparse, irregular, and also noisy. However, many imputation methods have been investigated and implemented in the learning framework for performance improvement. Most of the existing methods have a drawback that they treat all features of physiological time series as a single multivariate input stream without considering the correlations between organ

¹<https://myhealthrecord.gov.au>

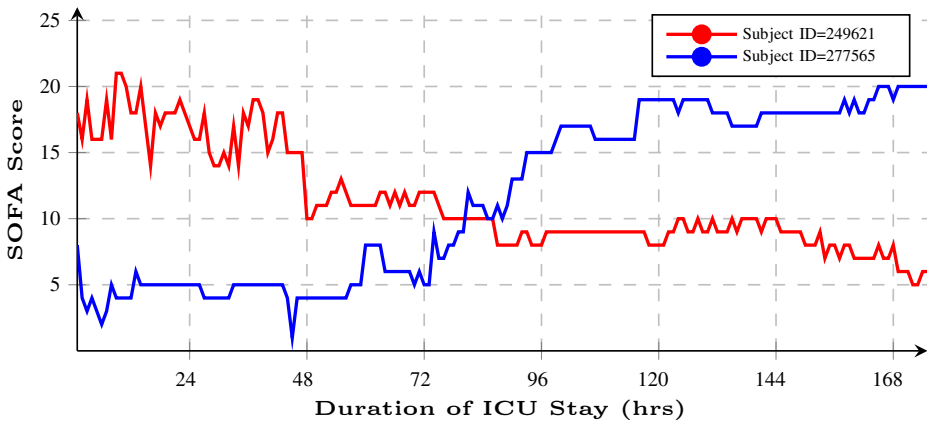


Figure 1 Two different SOFA trajectories of ICU patients. The patient (ID=249621, red) was initially in a critical condition (measured using SOFA score), but gradually improved and eventually discharged. In contrast, the conditions of the patient (ID=277565, blue) deteriorated in ICU and finally passed away

systems in a human body (e.g., heartbeats are correlated to blood pressure). As a result, this unfortunate ignorance can be detrimental to the performance of a prediction model. Consequently, the existing works are unavoidably biased and sensitive to a few physiological features in their learning process and have ignored the abnormal organ functions that may have a significant contribution to the prediction model in training.

1.3 Solution

To address the above-mentioned problems, we developed a novel deep learning model, namely, Attention Multi-task Recurrent Neural Networks (AMRNN) for continuously analysing multiple organ systems. In order to learn the multivariate time series of physiological features of each organ system separately, the approach of Multi-task LSTMs is used. Furthermore, a model of shared LSTM Layer is applied to utilise the temporal and functional relationships between organ systems to achieve better prediction results. Our model demonstrated a promising new solution for learning the correlated feature in multivariate time series with interpretable results.

1.4 Contributions

In summary, Our contributions are threefold:

- A novel multi-task LSTMs framework is proposed to learn each of the individual human organ systems simultaneously. The Attention mechanism was applied to capture the temporal and functional correlations between systems. As a result, better prediction performance is obtained.
- To learn correlated temporal features, the shared LSTM layer is applied to capture the important correlated features overtime by integrating both task-specific and cross-task interactions through the sequential EHR. This has suggested a new way of processing correlated temporal features.

- We have compared our method with state-of-the-art methods on real-world data with respect to the continuous illness severity prediction. The experimental results have shown that our framework outperforms all other methods.

2 Related work

EHR mainly consists of multivariate time series with various types of medical variables. In practice, time-series deep models have been widely adopted for analytic tasks, such as mortality risk estimation, phenotyping analysis, and disease modelling.

2.1 Multivariate time-series deep modelling

To address the above analytic challenges, many sophisticated models were developed over the pasting years. Wang et al. [27] investigate disease-specific features embedded in descriptive data for analysing the phenotype of patients. To improve the model performance, Zhou et al. [31] and Nie et al. [18] take the consistency of multiple modalities and the task-specific features into consideration for performance improvement. Unfortunately, the performance of the above prediction model is impacted by not utilising the temporal information embedded in the multivariate time series data. In particular, multivariate time series features can properly reflect the illness severity of a patient. To utilising the temporal information embedded in time series data, Pham et al. [20] adopted RNNs models to learn the time series feature to estimate ICU patients' illness severity. For performance improvement, Chen et al. [6], Lipton et al. [16] and Chen et al. [5] consider time interval while integrating RNNs for investigating irregular EHR time series data. However, the above works process the temporal feature in a heuristic manner by using a monotonically decreasing function. Therefore, these frameworks may cause under-parameterisation or over-parameterisation.

2.2 Multi-task learning

Multi-task Learning aims to leverage useful information contained in related learning tasks to improve the model performance, and it is widely used in computer vision [28, 29]. Abdalnabi et al. [1] proposed a multi-task CNN for attribute prediction, in which multiple CNNs tasks are used for learning image segments features, to unitise useful information a common layer is designed for determining the spatial correlation between image features. In spite of multi-task CNN, Chen et al. [6] proposed multi-task RNNs frameworks to exploit the correlations between sub-tasks by jointly learning EEG signals for intention recognition tasks. For multivariate time-series data analysis, Chen et al. [5] proposed a multi-task learning method to capture the temporal correlations embedded in learning tasks. The core design of multi-task analytics frameworks is to capture the intrinsic relationships between different learning tasks, and utilise the common features significant for performance improvement, Zhou et al. [30]. Harutyunyan et al. [10] extend the success of the heterogeneous model for time series learning. However, the performance of these models is impacted by over-parameterisation.

2.3 Sequence modeling

Despite recurrent neural networks(RNNs), many none-time series models are used in sequence modelling problem by adopting recurrent architecture, recently. Gehring et al. [8]

using convolution neural network with attention mechanisms, achieved a competitive result in sequence mining task. Extending the basic encoder-decoder architecture and attention mechanism Rocktäschel et al. [22] and Verga et al. [25] proposed a multi-head self-attention mechanism, and it demonstrated satisfactory performance in NLP tasks. In this paper, we apply and incorporate attention-based techniques in our proposed memory fusion network to effectively learn particular features across tasks.

3 Proposed method

In this section, we present the details of our novel framework based on multi-task deep RNNs with the attention mechanism to predict one of the commonly used illness severity scores, i.e., Sequential Organ Failure Assessment (SOFA) score, in ICU. Firstly, data preprocessing detail, including the cohort selection, data extraction, data cleansing, and feature extraction, is discussed. Secondly, The architecture detail of the proposed phased multi-task model is reviewed, showing the model is not only able to learn distinctive features from different human organ systems in time-series EHRs but also can exploit the temporal correlations between organ systems. Finally, we demonstrate how the Memory Fusion Network, which focuses on recognizing selectively features, is embedded in the proposed multi-task model to capture descriptive representations for result improvement.

3.1 Data preprocessing

We adopted the MIMIC-III V1.3 [12], which contains 53,423 de-identified adult patients from *Beth Israel Deaconess Medical Center* from 2001 to 2012 in this work. Following the convention, we exclude all patients who are younger than 16 (age < 16) and stay in ICU less than 24 hours. In this work, we consider each ICU stay as an independent data observation in our benchmark dataset, which eventually has 45,321 records. Figure 2 illustrates the detail data distribution regarding gender and age-groups.

To collect multi-variate physiological features, we have extracted 41 features, as shown in Table 1, with respect to different human organ systems from multiple tables in MIMIC III. More details of the extracted feature sets can be found in the supplementary material. The values of each feature within a time window will be averaged as the new value in that time slot. We have also considered three different time-window different lengths, including

Figure 2 Age distribution in the selected cohort

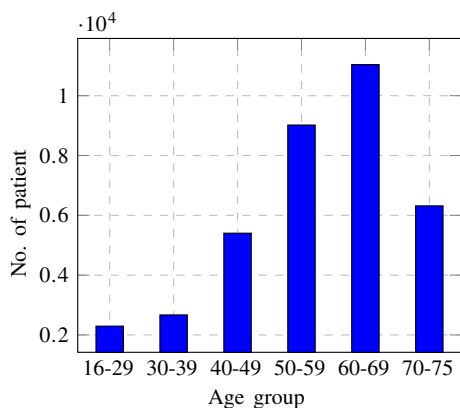


Table 1 List of 41 features extracted from MIMIC III

| Index | Feature name | Index | Feature name |
|-------|--------------------|-------|------------------------------------|
| 1 | BPd | 22 | HR |
| 2 | BPm | 23 | INR |
| 3 | BPs | 24 | Magnesium |
| 4 | Albumin | 25 | Mean Airway Pressure |
| 5 | ALT | 26 | O2 Flow |
| 6 | ApH | 27 | PaCO2 |
| 7 | Phosphorous | 28 | PaO2 |
| 8 | Alkaline Phosphate | 29 | Platelets |
| 9 | Bilirubin | 30 | Potassium |
| 10 | BUN | 31 | PT |
| 11 | Calcium | 32 | PTT |
| 12 | Calcium-Ionized | 33 | RR |
| 13 | Chloride | 34 | SaO2 (Arterial O2 Saturation) |
| 14 | Creatinne | 35 | Sodium |
| 15 | CVP | 36 | SpO2 (O2 saturation pulseoxymetry) |
| 16 | FiO2 | 37 | TCO2 |
| 17 | GCS | 38 | Temperature |
| 18 | Glucose | 39 | Urine |
| 19 | HCO3 | 40 | WBC |
| 20 | HCT | 41 | Weight |
| 21 | Hemoglobin | | |

1-hour, 3-hour, and 6-hour. For each stay record, all extracted features will be converted into a matrix with a variable number of rows as Figure 4 illustrated. 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, $i = 1, \dots, n$. In this way, the data samples can be represented by $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in R^{t_i \times D}$ (Figure 3). As pointed out in [21], extracted data are in a low quality due to missing values, irregular sampling, the outlier, etc. We have borrowed the same procedures in [21] to improve the data quality. For missing values of the d -th variable at t , we adopted the *forward-fill* imputation strategy in [16] as follows:

- If there is at least one valid observation at time t' , where $t' < t$, then $x_{t,d} := x_{t',d}$.
- If there are no previous observations, then the missing value will be replaced by the median value over all measurements.

This strategy is inspired by the fact that measurements are recorded at intervals proportional to the rate at which the values are believed or observed to change (Figure 4).

3.2 Multi-task recurrent neural networks

The recurrent neural networks (RNNs) [7] is capable to process arbitrary sequential inputs by applying a transaction function to its *hidden vector* \mathbf{h}_t recursively. However, RNNs have difficulties learning long-range dependencies overtime. The components of the gradient vector will vanish or explode exponentially over a very long sequence. In order to address

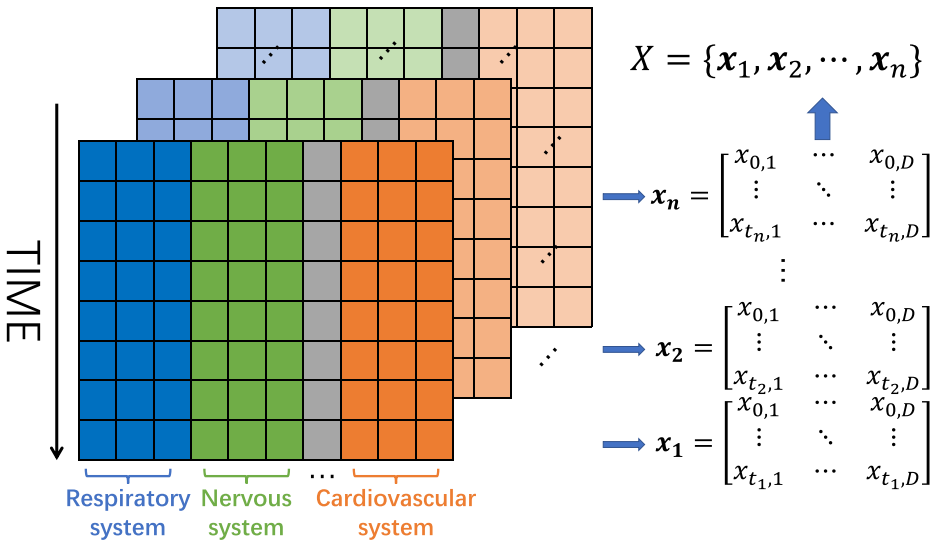


Figure 3 The structure of data organization

the vanishing problem, a LSTM (Long Short-Term Memory) network [11] has been proposed by implementing gating functions into RNNs. By comparing to RNNs, at each time step the LSTM maintains a hidden vector and a memory vectors for controlling state updates and outputs in [9]. The LSTM unit consists of i, f, o, c , which are the input gate, forget gate, output gate, memory cell. The forget gate is used to control the amount of memory to be “forgotten” in each unit, while the input gate controls the update of each time step and the output gate rules the exposure of memory state of each time step. The activation function of LSTM can be computed as followings:

$$h_t = \begin{cases} 0 & t = 0 \\ f(h_{t-1}, x_t) & \text{otherwise,} \end{cases} \tag{1}$$

where x_t is the input of current time-step, and h_{t-1} is the hidden state of previous time-step.

The LSTM transition equations are defined as follows:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\ c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{2}$$

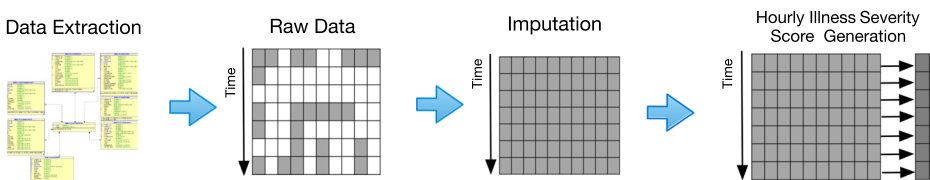


Figure 4 The work-flow of data preprocessing

where x_t is the input at a time t , W are weights, bs are bias terms, and σ denotes the logistic sigmoid function.

In order to capture the correlations organ systems, we have used a shared layer to exploit temporal correlations between different systems. Figure 5 has shown the structure of our MTRNN. Specifically, features from different systems are fed into different learning task. For example, features denoted by x from respiratory system and the ones denoted by 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 tasks and aims to capture intrinsic features in long-short terms, respectively. As human organs collaboratively work together, it is believed that there must be correlations between organ systems, which can be beneficial to learning tasks. To capture such kind of temporal correlation between systems, we added a shared layer $LSTM^{(s)}$, as shown in the middle of Figure 5, in our framework. The shared hidden layer fully connects with all the other task-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 (2). In contrast, we have modified the activation function for each LSTM $h_t^{(m)}$, which learns different organ features as below:

$$\mathbf{h}_t^{(m)} = \begin{cases} 0 & t = 0 \\ f\left(\mathbf{h}_{t-1}^{(m)} \odot \mathbf{h}_{t-1}^{(s)}, x_t^{m,i}\right) & \text{otherwise,} \end{cases} \tag{3}$$

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

$$c_t^{(m)} = f_t c_{t-1}^{(m)} + i_t^{(m)} \tanh(W_{xc} x_t^{m,i} + W_{hc} h_{t-1}^{(m)} \odot h_{t-1}^{(s)} + b_c^{(m)}), \tag{4}$$

where $x_t^{m,i}$ is the input at time t . $h_{t-1}^{(m)}$ is the output of (3) when $t - 1$. The shared hidden layer outputs $h_{t-1}^{(s)}$ when $t - 1$.

3.3 Attention mechanism

Attention mechanisms as been shown to produce state-of-the-art results in computer vision and natural language task. When combining with sequential learning models e.g. RNNs,

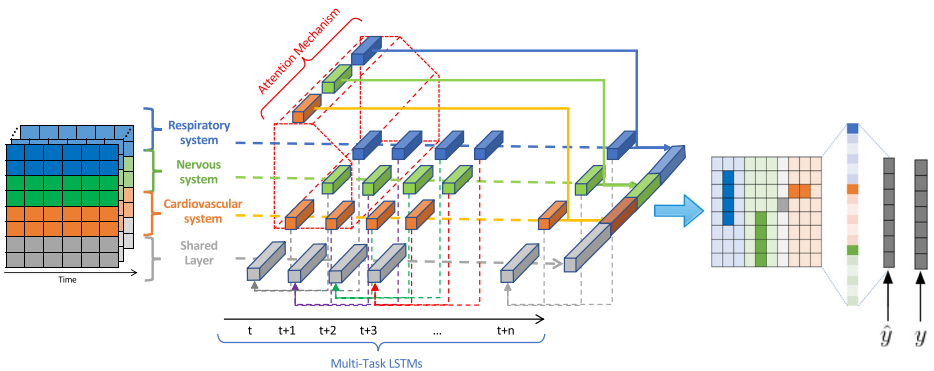


Figure 5 Demonstration of our proposed Multi-task LSTM Architecture with a shared hidden layer. Note that this structure has not detailed structure of the attention layer

attention filters the perceptions that can be stroed in memory while perceiving the surrounding information, and then adjusting the focal point over time. In order to taking the advantages of attention, we incorporate our multi-task RNNs with the attention As shown in Figure 6. So that our model can pay more “attention” selective important feature overtime. For each patient, the attention weight is calculated using the dot-product \odot , of the hidden state for every feature in the input. Therefore the score for the t -th feature $score_t$ is calculated as follows:

$$score_t = h_f^\top \hat{h}_s \tag{5}$$

where \hat{h}_s is the concatenated hidden state of LSTM, in which the t -th the input, and h_f is the learned feature of the input.

The weight of the t -th input W_t can be computed by using the out put of $score_t$ as follows:

$$W_t = \frac{\exp(score_t)}{\sum_{t'} \exp(score_{t'})} \tag{6}$$

here, t' means the all input features.

The final output a_f can be computed by using he weight and the hidden state of the feature as a convex sum of hidden states h_t :

$$a_f = \sum W_t h_t \tag{7}$$

The structure of attention mechanisms, in which the attentional decisions are made independently, is illustrated in Figure 6. The reasons of adopting the attention mechanisms are two-fold:

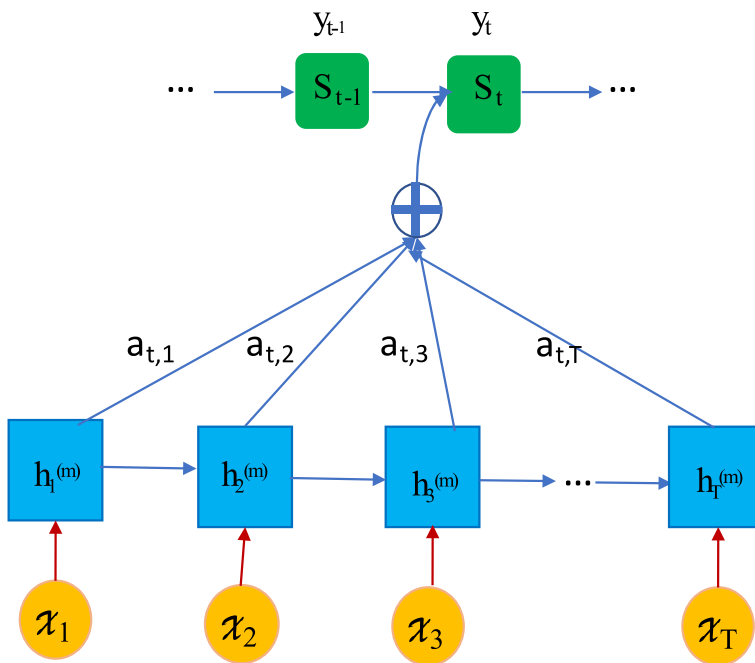


Figure 6 Illustration of the structure of dot-product based attention mechanism

- the attention function gives high weight to the feature that strongly affect the vector representation of the whole input
- it establishes direct short-cut connections between the target and the source.

3.4 Loss function

In this setting, we use a softmax classifier to predict the label \hat{y} from a discrete set of classes Y for ICU each stay X . The classifier takes the output of attention vector \hat{h}_t as input:

$$\hat{p}(y | X) = \text{softmax}(W^{(X)}\hat{h}_t + b^{(X)}) \quad (8)$$

$$\hat{y} = \arg \max_y \hat{p}(y | X) \quad (9)$$

The loss function is the negative log-likelihood of the true class labels \hat{y} . The loss function is defined as follows:

$$L = -w \sum_{i=1}^c t_i \log t_i + \lambda \|y - \hat{y}\|_F^2 \quad (10)$$

where $t \in \mathfrak{N}^m$ is the one-hot representation of groundtruth, $y \in \mathfrak{N}^m$ is the estimated probability for each class by Softmax, C is the number of output classes, and the estimate value \hat{y} , and λ is hyper-parameter of $L2$ -norm regularisation.

To alleviate over-fitting, we coupled Dropout with $L2$ -norm. Dropout prevents co-adaptation of hidden units by randomly omitting feature detectors from the network. The $L2$ -norm imposes an additional constraint over the weight vectors by rescaling w to have $\|w\| = x$. The training details will be further introduced in Section 4.

3.5 Complexity

In this section, we discuss the asymptotic complexity of our framework and how it offers a higher degree of parallelism than the other single LSTM based framework. We assume that all hidden dimensions are d and n is the sequence length.

Since our proposed multi-task LSTM is implemented in parallel fashion, so we only focus on the most complex task which is the Shared LSTM Layer in our framework, and its time complexity is $O(nd^2)$ [19]. In addition, the complexity of dot-product based attention mechanism is $O(n^2d)$ [24] Thus the overall complexity is $O(nd^2 + n^2d)$. However, note that since $n \gg d$, i.e., the hidden dimension d in our case is far greater than the sequence length n . The overall complexity of our framework can be simplified as $O(nd^2)$. In sum, the complexity of our model is identical to a basic LSTM model.

4 Experiments

4.1 Data description and experiment design

we conducted extensive experiments to evaluate the performance of the proposed model on a publicly available benchmark dataset *MIMIC III*.² We compare our model against state-of-the-art algorithms and several baselines. Meanwhile, we also investigate the influence of the multi-task structure and the Memory Fusion Network via experiments.

²<https://mimic.physionet.org/>

4.2 Details of dataset and settings

We have investigated all methods on the version 1.3 MIMIC-III dataset [12], which is publicly available. In this work, we focus only on adult ICU patients who are older than and equal to 16 years of age and has records for more than 24 hours. We treat each admission as an independent data sample. We randomly select 80% of 45,321 ICU stays as training data, while another 20% of data are used as testing and validation. To select best parameters, we have employed a 10-fold cross-validation schema in the experiments. All the experiments are repeatedly run 10 times.

All neural networks were implemented with the TensorFlow and Keras frameworks and trained on 2 Nvidia 1080 Ti GPUs from scratch in a fully-supervised manner. To minimise the cross-entropy loss, we employed the stochastic gradient descent with Adam update rule [13]. The network parameter is optimised with a learning rate of 10^{-4} . The keep probability of the dropout operation is 0.5. The number of neurons in the input and output layers in the AMRNN model is fixed at 41. and the λ is 4×10^{-4} .

4.3 Comparison methods

The effectiveness of AMRNN is evaluated using ROC, AUC, Precision, Recall, and F1-Score by comparing with the following state-of-the-art algorithms and baselines methods.

1. **GRU-ATT**: Nguyen et al. [17] proposed a GRU-based (Gated Recurrent Unit) attention networks for mortality risk estimation.
2. **HMT-RNN**: Harutyunyan et al. [10] have employed RNNs (recurrent neural networks) for the prediction of in-hospital mortality.
3. **pRNN**: Aczon et al. [2] take encounter records (physiologic feature, laboratory test, and administered drugs) into consideration while using an RNN-based framework for mortality prediction.
4. **RNN**: A standard recurrent neural network (RNNs) is implemented as one of the baselines.
5. **MTRNN**: A standard multi-task RNNs without attention network is implemented as one of the baselines.
6. **RNNATT**: A standard single task RNNs with attention network is implemented as another baseline method.

Apart from a set of state-of-the-art methods, we comparing the proposed method against some representative classification baseline methods, including **Support Vector Machines** (SVMs), **Decision Tree** (DT), **Linear Discriminant Analysis** (LDA), **Random Forest** (RF), and **XGboost**. All the parameters have been fine-tuned using a Grid-search scheme and the best results with the optimal parameters are reported.

The SOFA score is useful while envisioning the developing of critically sick patients, the mortality risk estimation is based on the highest SOFA score during a patient's ICU stay as shown in Table 2. We follow the class setting in [5] categorised by critical care experts.

Table 2 The SOFA scores and its corresponding label

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

4.4 Evaluation metrics

To choose appropriate evaluation metrics in this study, we applied precision and F1, which have been widely used in this field of studies, to evaluate accuracy. These metrics have been adapted to evaluate the accuracy of a set of correct predicted and are defined as follows:

$$Accuracy = \frac{TP+TN}{Total} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (14)$$

4.5 Discussion

To consider the overall classification performance, we have reported the results of all the methods measured by Accuracy in Table 3. It is clear that our approach performs better than all compared methods. Particularly, the proposed multi-task framework (#12) performs much better than most of the counterparts with the Gated Recurrent Unit with attention mechanism GRU-ATT [17] (#1: 83.05%) and RNNATT (#11: 83.77%) settings. It may contribute to the exploitation of temporal correlations between different human organ systems by the Memory Fusion Network.

To evaluate the impact of modelling missingness and the effectiveness of data imputation strategy. We conducted experiments on the raw data and processed data. As illustrated in Table 4, the imputation strategy is effectively improving the data quality, where the performance of all classification models are all improved. In addition, our model can effectively handle missing values in multivariate time-series data and achieved best the result.

Table 3 Overall performance comparison (Accuracy)

| 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 | 0.8041 |
| 10 | MTRNN | 0.8330 |
| 11 | RNNATT | 0.8377 |
| 12 | AMRNN | 0.8742 |

Table 4 Classification performance on Raw ICU data, data with missing values, and imputed data. Model performances measured by Accuracy

| Index | Method | Raw data | Processed data |
|-------|---------|----------|----------------|
| 1 | SVM | 0.4103 | 0.6983 |
| 2 | RF | 0.3981 | 0.7153 |
| 3 | DT | 0.3915 | 0.7230 |
| 4 | LDA | 0.3851 | 0.7122 |
| 5 | XGBoost | 0.4463 | 0.6334 |
| 6 | MTRNN | 0.5717 | 0.8330 |
| 7 | AMRNN | 0.6332 | 0.8742 |

To evaluate the model performance with respect to different sizes of training dataset, we randomly sub-sample three smaller datasets of 30%, 60%, and 90% admissions from the entire experimental dataset while keeping the same class distribution. We compare our proposed model with all the baseline methods and the second and third best models in overall performance test, i.e. **GRU-ATT**, and **HMT-RNN** on 1-hour time-window dataset. From Figure 10, It can be observed that all the methods achieve better performance while more training samples are given. However, the prediction performance improvements of baselines are limited by comparing to deep learning methods. The proposed model achieves the best performance on all sub-sample datasets and the performance gap between **AMRNN** and baselines will show continuing growth when more data become available (Figure 7). The Receiver Operating Characteristic (ROC) curve can demonstrate the discrimination capability of a classifier by plotting the True Positive Rate against the False Positive Rate in a range of threshold values. In Figure 8, we noted that the ROC curves of all the categories are very far from the 45-degree diagonal and close to the upper left corner of the ROC space. The areas under each of these six ROC curves (AUC) are shown in Table 5, and the average value is about **96.72%** showing an excellent performance. Also, we can observe that the proposed method is very sensitive to Class 1 and Class 6, which are two critical scenarios in the ICU. In other words, the proposed method can not only effectively recognise the

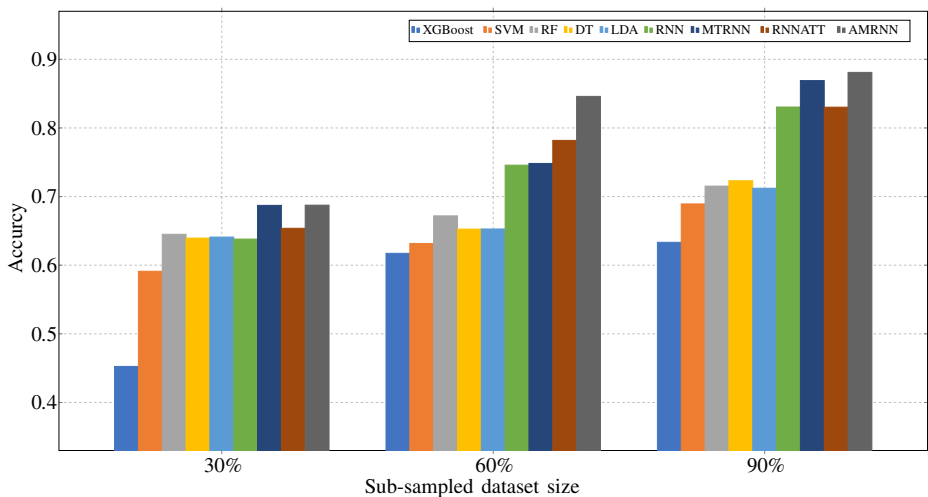


Figure 7 Acc. vs Data Proportion: Prediction accuracy with different training set sizes. x-axis = sub-sampled dataset size; y-axis = accuracy

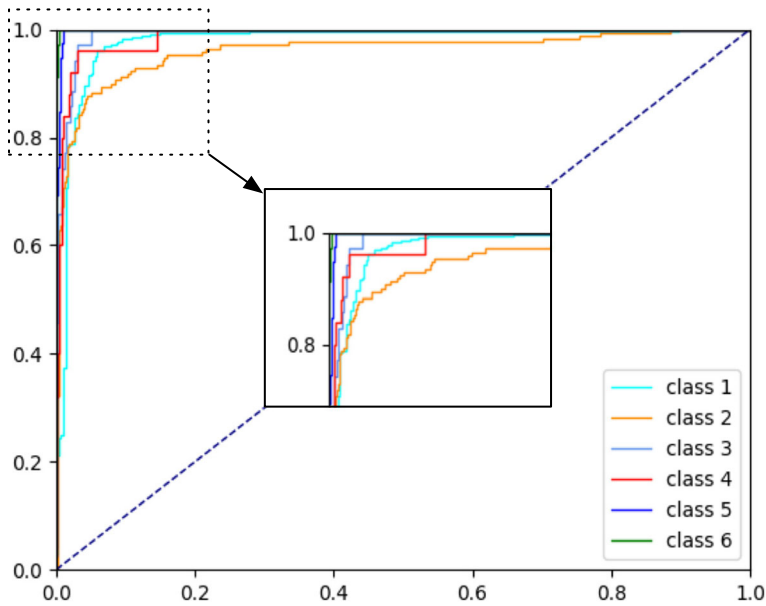


Figure 8 The ROC curves showing the discrimination capability of a classifier. x-axis, True positive size; y-axis, False positive

critical conditions of ICU patients but is also reasonably good at discrimination of intermediate conditions of patients.

The window size is another important parameter that impacts on the classification performance. To evaluate the influences of window size, we have evaluated the algorithms with respect to different sizes (1-hour, 3-hour, and 6-hour) and report the performance results

Table 5 Evaluation on the influence of Memory Fusion Networks and Multi-Task Model

| | Method | Class1 | Class2 | Class3 | Class4 | Class5 | Class6 | Average |
|----------|--------|--------|--------|--------|--------|--------|--------|---------|
| Percison | RNN | 0.8821 | 0.6707 | 0.4074 | 0.5714 | 0.6136 | 0.6356 | 0.7893 |
| | MTRNN | 0.8857 | 0.6535 | 0.6470 | 0.6250 | 0.7326 | 0.7241 | 0.8219 |
| | RNNATT | 0.8921 | 0.7021 | 0.6667 | 0.6250 | 0.7826 | 0.7241 | 0.8219 |
| | AMRNN | 0.9391 | 0.7393 | 0.6728 | 0.6538 | 0.8076 | 0.9062 | 0.8698 |
| Recall | RNN | 0.9454 | 0.4823 | 0.5314 | 0.3612 | 0.6923 | 0.4705 | 0.8040 |
| | MTRNN | 0.9569 | 0.5882 | 0.6285 | 0.6428 | 0.5815 | 0.6174 | 0.8330 |
| | RNNATT | 0.9540 | 0.5941 | 0.6285 | 0.5210 | 0.6438 | 0.6458 | 0.8377 |
| | AMRNN | 0.9625 | 0.7393 | 0.6428 | 0.6538 | 0.8076 | 0.9062 | 0.8742 |
| F1 | RNN | 0.9127 | 0.5377 | 0.5348 | 0.3670 | 0.5606 | 0.5612 | 0.7936 |
| | MTRNN | 0.9200 | 0.6191 | 0.6376 | 0.3681 | 0.5806 | 0.6667 | 0.8217 |
| | RNNATT | 0.9223 | 0.6412 | 0.6945 | 0.5652 | 0.6268 | 0.7581 | 0.8343 |
| | AMRNN | 0.9394 | 0.7283 | 0.7714 | 0.6777 | 0.6465 | 0.8788 | 0.8703 |
| AUC | RNN | 0.9105 | 0.8411 | 0.9168 | 0.9360 | 0.9780 | 0.9815 | 0.9273 |
| | MTRNN | 0.9358 | 0.8772 | 0.9454 | 0.9294 | 0.9597 | 0.9841 | 0.9386 |
| | RNNATT | 0.9281 | 0.8846 | 0.9635 | 0.9614 | 0.9795 | 0.9318 | 0.9414 |
| | AMRNN | 0.9596 | 0.9249 | 0.9647 | 0.9652 | 0.9819 | 0.9980 | 0.9657 |

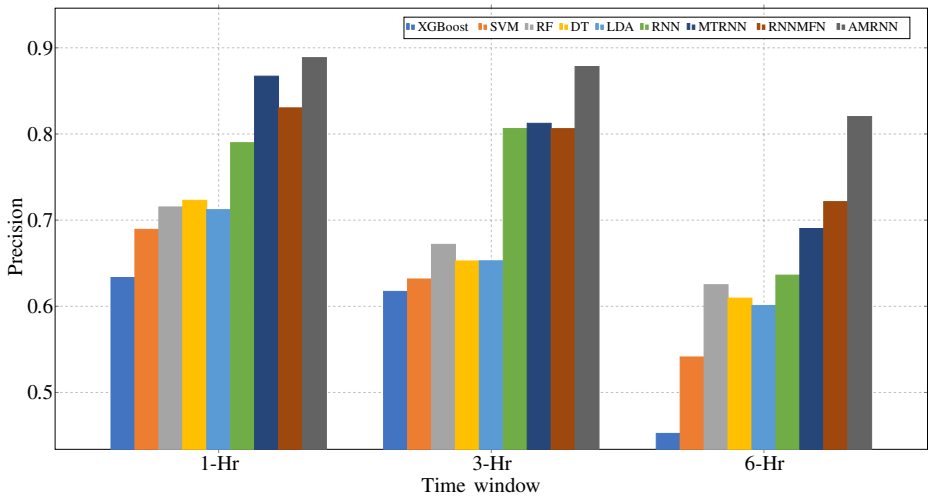


Figure 9 Precision vs Time window: Precision on different length of prediction time windows. x-axis = time window; y-axis = precision

in Figures 9 and 10. The experiment results are measured by **Precision** and **Recall**. The two figures, it clearly shows that our method achieves the best performance, over the baseline methods. Also, we have observed that prediction performance drops slightly with the increase of time windows. This may be because the variations of the medical condition can change dramatically, for better or worse, over a longer period of time.

To investigate the influence of Memory Fusion Networks and Phased model, we have built up three baseline models and reported the results in Table 5, which illustrates the classification performance with respect to each class. The performance measurements we

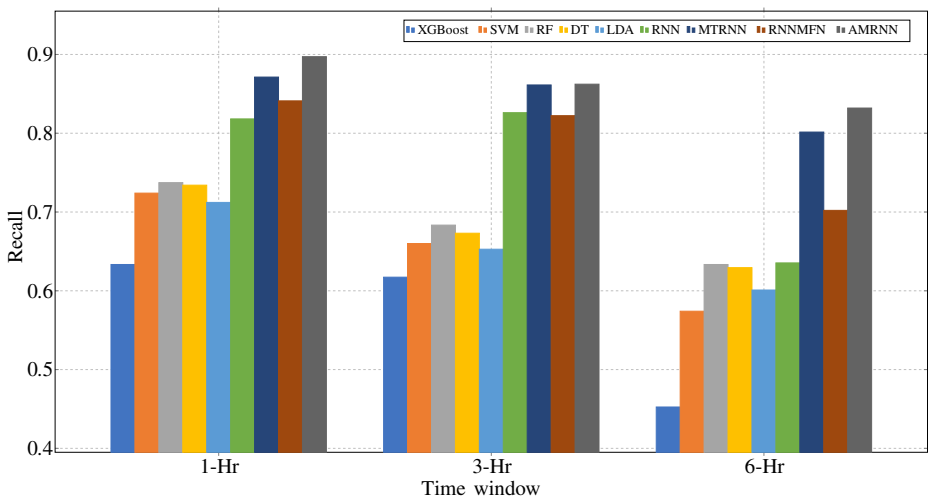


Figure 10 Recall vs Time window: Recall on different length of prediction time windows. x-axis = time window; y-axis = precision

have used use include precision, recall, F1 score, AUC, and test error. For individual class, the criteria is calculated with a one-versus-all method. We observe consistently superior performance of our method against all the baseline models for all criterion and all individual classification tasks. From the results in Table 5, our method achieves a better result in all cases than all comparison methods.

5 Conclusion

In this paper, we propose a novel deep learning framework that simultaneously analyses different human organ systems to predict illness severity of patients in the ICU. our framework based on multi-tasks LSTMs and it treat each organ system separately and also exploit the correlations between organ systems by a shared unit. To our best knowledge, this work is the first to analyse ICU patient systematically. To deal with problems raised by data quality, we have applied attention mechanisms to gives high weight to the “important” feature of the input to further improve the model performance. Through the comprehensive experiments, we have shown that our approach outperforms all the compared methods and baselines in the scenario of illness severity prediction, which is actually a multi-class problem.

6 Future work

In our future work, for missing value imputation, we intend to incorporate a mask of missing data to indicate the placement of imputation values or missing values. So that the model can not only captures the long-term temporal dependencies of time-series observations but also utilizes the missing patterns to further improve the prediction results. In addition, we plan to investigate the most sophisticated sharing mechanisms in the RNNs based multi-task architecture to enhance the feature representation.

References

1. Abdalnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task cnn model for attribute prediction. *IEEE Trans. Multimedia* **17**(11), 1949–1959 (2015)
2. Aczon, M., Ledbetter, D., Ho, L., Gunny, A., Flynn, A., Williams, J., Wetzel, R.: Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. [arXiv:1701.06675](https://arxiv.org/abs/1701.06675) (2017)
3. Binder, H., Blettner, M.: Big data in medical science—a biostatistical view: Part 21 of a series on evaluation of scientific publications. *Dtsch. Arztebl. Int.* **112**(9), 137 (2015)
4. Bouch, D.C., Thompson, J.P.: Severity scoring systems in the critically ill. *Continuing Education in Anaesthesia. Critical Care & Pain* **8**(5), 181–185 (2008)
5. Chen, W., Wang, S., Long, G., Yao, L., Sheng, Q.Z., Li, X.: Dynamic illness severity prediction via multi-task rnns for intensive care unit. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 917–922. IEEE (2018)
6. Chen, W., Wang, S., Zhang, X., Yao, L., Yue, L., Qian, B., Li, X.: Eeg-based motion intention recognition via multi-task rnns. In: Proceedings of the 2018 SIAM International Conference on Data Mining, pp. 279–287. SIAM (2018)
7. Elman, J.L.: Finding structure in time. *Cogn. Sci.* **14**(2), 179–211 (1990)
8. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1243–1252. JMLR. org (2017)
9. Graves, A.: Generating sequences with recurrent neural networks. [arXiv:1308.0850](https://arxiv.org/abs/1308.0850) (2013)

10. Harutyunyan, H., Khachatrian, H., Kale, D.C., Steeg, G.V., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. arXiv:1703.07771 (2017)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* (1997)
12. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
14. Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A., et al.: The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest* **100**(6), 1619–1636 (1991)
15. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama* **270**(24), 2957–2963 (1993)
16. Lipton, Z.C., Kale, D.C., Wetzell, R.: Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare* (2016)
17. Nguyen, P., Tran, T., Venkatesh, S.: Deep learning to attend to risk in icu. arXiv:1707.05010 (2017)
18. Nie, L., Zhang, L., Yang, Y., Wang, M., Hong, R., Chua, T.S.: Beyond doctors: Future health prediction from multimedia and multimodal observations. In: *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 591–600. ACM (2015)
19. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. arXiv:1606.01933 (2016)
20. Pham, T., Tran, T., Phung, D., Venkatesh, S.: Deepcare: a deep dynamic memory model for predictive medicine. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 30–41. Springer (2016)
21. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare mimic datasets. arXiv:1710.08531 (2017)
22. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiský, T., Blunsom, P.: Reasoning about entailment with neural attention. arXiv:1509.06664 (2015)
23. Shann, F., Pearson, G., Slater, A., Wilkinson, K.: Paediatric index of mortality (pim): a mortality prediction model for children in intensive care. *Intensive Care Med.* **23**(2), 201–207 (1997)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
25. Verga, P., Strubell, E., McCallum, A.: Simultaneously self-attending to all mentions for full-abstract biological relation extraction. arXiv:1802.10569 (2018)
26. Vincent, J., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., Thijs, L.: The sofa score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Med.* **22**(7), 707–710 (1996)
27. Wang, S., Chang, X., Li, X., Long, G., Yao, L., Sheng, Q.Z.: Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE Trans. Knowl. Data Eng.* **28**(12), 3191–3202 (2016)
28. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 676–684 (2015)
29. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via structured multi-task sparse learning. *Int. J. Comput. Vis.* **101**(2), 367–383 (2013)
30. Zhou, J., Yuan, L., Liu, J., Ye, J.: A multi-task learning formulation for predicting disease progression. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 814–822. ACM (2011)
31. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via fused sparse group lasso. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1095–1103. ACM (2012)

Affiliations

Weitong Chen¹ · Guodong Long¹ · Lina Yao² · Quan Z. Sheng³

Guodong Long
guodong.long@uts.edu.au

Lina Yao
lina.yao@unsw.edu.au

Quan Z. Sheng
michael.sheng@mq.edu.au

¹ The University of Queensland, St Lucia, Australia

² The University of New South Wales, Sydney, Australia

³ Macquarie University, North Ryde, Australia