Deep Feature Learning for Disease Risk Assessment Using multi-task recurrent neural Networks for dynamic illness

S.Midhun
Research Scholar
Dept of Computer Science and Engineering
Annamalai University

Dr. A. SUHASINI
Professor
Dept of Computer Science and Engineering
Annamalai University

Dr.A.Subitha
Associate Professor
Dept of Computer Science and Engineering
St. Xaviers Catholic College of Engineering

ABSTRACT

Analytics on Intensive Care Unit data focus on mortality risk prediction & phenotyping analysis. Mainly they have drawbacks in providing evidence for decision making in a dynamically changing clinical environment. In this paper, we propose an 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.

Keywords: Deep learning; self-attention;disease prediction Multi-task learning Illness; severity prediction

1 Introduction

Artificial intelligence (AI) is defined as the technology that uses computer knowledge to represent intelligent behavior with nominal human involvement, and Artificial Intelligence (AI) is considered as a subset of AI techniques. Usually, this kind of intelligence is commonly acknowledged as having begun with the innovation of robotics [1]. With the rapid growth of electronic speeds and programming, computers may display intelligent behaviorsimilar to that of humans in the near future. This is because of the large advancements happening in contemporary ideas in the development of AI [2]. Artificial intelligence can be defined as human intelligence which is performed by machines. In computer science, it is defined as the machine's capacity to emulate intelligent behavior by itself, using nothing but AI [3]. The applications of AI in medicine are developing quickly. In 2016, AI projects coupled with medicine drew in more speculation from the global economy than other projects [4]. In medicine, AI refers to the utilization of automated diagnosis processes and the treatment of patients who require care. Increased AI utilization in prescription will allow a considerable amount of the role to be automated, opening up medicinal experts' time to be used in performing different obligations, ones that cannot be automated. As such, this technology promises progressively significant utilization in the field of human resources (HR).In general, AI is categorized as supervised (i.e., consists of output variables that are predicted from input variables) [5] or unsupervised (i.e., deals with clustering of different groups for a particular intervention). AI is used to determine complex models, and extract medical knowledge, exposing novel ideas to practitioners, and specialists [2]. In clinical practice, AI predictive models can highlight enhanced rules in the decision-making regarding individual patient care. These are also capable of autonomous diagnosis of different diseases under clinical regulations [4,6–8]. In [9], the incorporation of these models in drug prescription can save doctors and offer new medical opportunities in pathology identification. With AI models, it can also be possible to improve quality of medical data, reduce fluctuations in patient rates, and save in medical costs. Therefore, these models are frequently used to investigate diagnostic analysis when compared with other conventional methods [10]. To reduce the death rates caused by chronic diseases (CDs), early detection and effective treatments are the only solutions [11]. Therefore, most medical scientists are attracted to the new technologies of predictive models in disease forecasting [12]. These new advancements in medical care have been expanding the accessibility of electronic data and opening new doors for decision support and productivity improvements [13]. AI methods have been effectively utilized in the computerized interpretation of pneumonic capacity tests for the differential analysis of CDs. It is expected that the models with the highest accuracies could gain large importance in medical diagnosis. Due to the low-progress nature of CDs, it is important to make an early prediction and provide effective medication. Therefore, it is essential to propose a decision model which can help to diagnose chronic diseases and predict future patient outcomes. While there are many ways to approach this in the field of AI, the present study focuses distinctly on AI predictive models used in the diagnosis of CDs, which highlights the importance of this study. In this study, we conducted a systematic literature review of different state-of-art of predictive models, and our significant contribution in this paper is to develop comparative model analysis to propose model optimization. In comparison to the conventional data analysis techniques, this review article will able to find promising results that enhance the quality of patient data and analysis of specific items that are related to AI algorithms in medical care.

2. Methods

2.1. Search Strategy

The systematic literature search was conducted through the libraries of PubMed (Medline) and Cumulative Index to Nursing and Allied Health Literature (CINAHL). Keywords like ‘chronic diseases’, ‘predictive models’, ‘ML in CD diagnosis’, and ‘model classifiers’ were used during the document search. The search was conducted in January 2020 and resulted in 453 documents. The documents were filtered based on its publication dates ranging from 2015 to 2019 to evaluate the latest literature on ML classifiers in CD prediction.

2.2. Selection Criteria

The title and abstract of the individual articles were retrieved based on the mentioned search terms. Finally, a few of the items were found to be eligible to fulfill the research objectives. This research only describes predictive models used to perform CD diagnosis and does not concentrate on overall trends in AI medicine. Further article revision was conducted to filter the duplicates between the two databases. Moreover, the inclusion and exclusion criteria of our review were based on time, methodological quality and language. Reports and other studies published before 2015 were excluded as outside the limitations on the timeframe of this study. The inclusion criteria used in Pub Med and CINAHL are as follows: free full text, English, original papers and research outcomes. We excluded 276 items among the total search documents because of duplication. The remaining 177 were screened to match the methodologies related to the current research topic.

2.3. Data Extraction

Data evaluation was conducted in two phases. In the first phase, depending on the inclusion criteria, 55 documents were identified for extended revisions. In the second phase, two individual researchers (GB and GGS) were equally distributed for quality check. As discussed, the proposal of a precise model in CD diagnosis was considered as the main focus of this paper. Therefore, articles were extracted based on the authors’ information, the study design of sampling pattern and method types, and diagnostic criteria. The analysis of each article was individually revised and recorded.

2.4. Quality Evaluation

Quality assessment check was accomplished by the adoption of the Newcastle–Ottawa Scale (NOS), which is a renowned method in the assessment of study relevance and research interest [14]. The quality of each published article was evaluated as weak (0–4), moderate (5–6), or strong (7–9). Each selected study score was recorded in separate excel sheets to compute whether an individual paper was suitable or not for this review. Ultimately, 22 studies were selected, which are in line with the predictive models in the CD diagnosis (Figure 1). Based on their content, the selected papers were tabled into predictive models used in CD identification (Table 1) and pathologies with model usage, along with their strengths and limitations (Table 2).

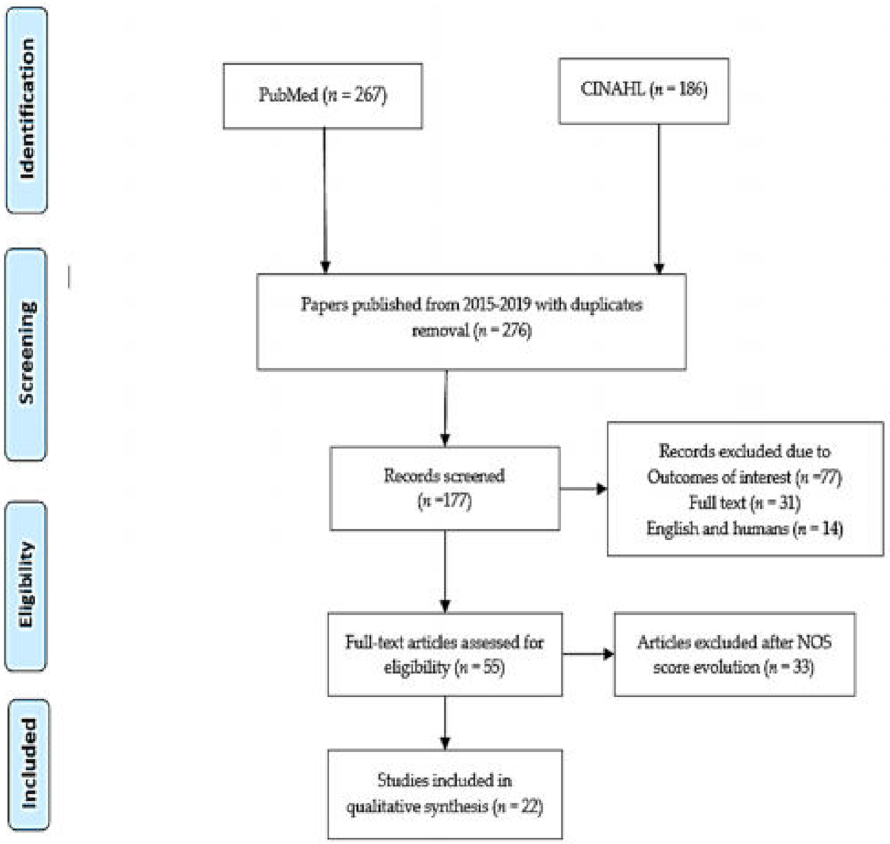
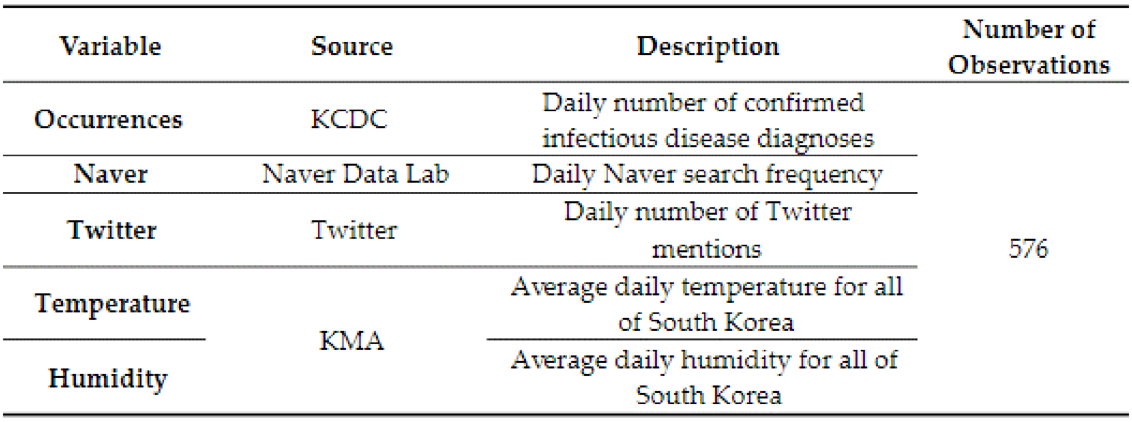

Identification
Screening
Eligibility
Included
PubMed (n = 267)
CINAHL (n = 186)
Papers published from 2015-2019 with duplicates removal (n = 276)
Records screened (n =177)
Records excluded due to
Outcomes of interest (n =77)
Full text (n = 31)
English and humans (n = 14)
Full-text articles assessed for eligibility (n = 55)
Articles excluded after NOS score evolution (n = 33)
Studies included in qualitative synthesis (n = 22)

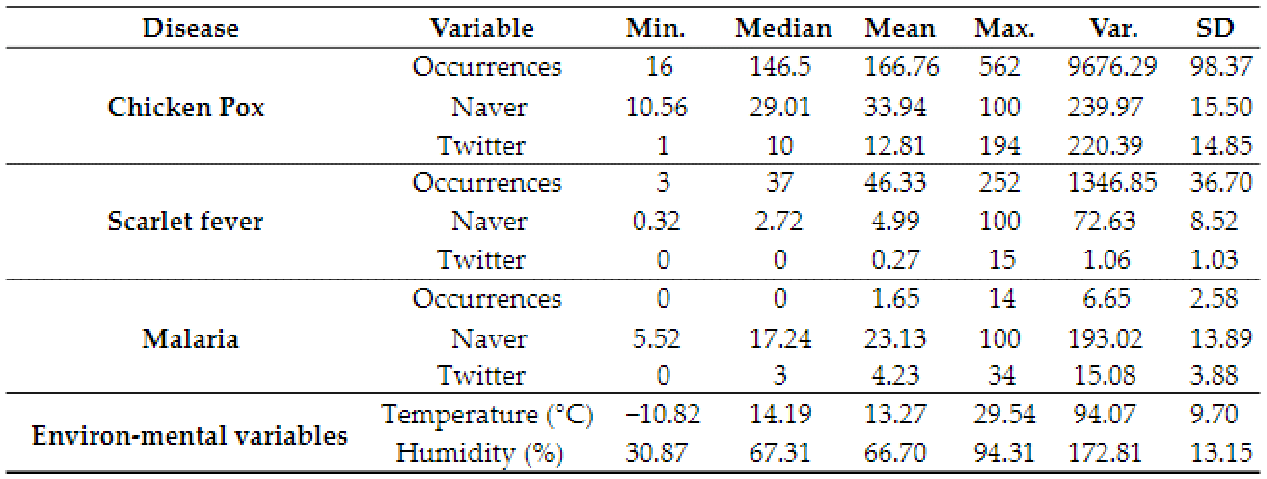
Figure 1. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) diagram [15].

Table 1. Description of data types. KCDC: Korea Center for Disease Control; KMA: Korea Meteorological Administration's weather information open portal.



Variable	Source	Description	Number of Observations
Occurrences	KCDC	Daily number of confirmed infectious disease diagnoses	576
Naver	Naver Data Lab	Daily Naver search frequency	
Twitter	Twitter	Daily number of Twitter mentions	
Temperature	KMA	Average daily temperature for all of South Korea	
Humidity		Average daily humidity for all of South Korea	

Table 2. Data statistics.



Disease	Variable	Min.	Median	Mean	Max.	Var.	SD
Chicken Pox	Occurrences	16	146.5	166.76	562	9676.29	98.37
	Naver	10.56	29.01	33.94	100	239.97	15.50
	Twitter	1	10	12.81	194	220.39	14.85
Scarlet fever	Occurrences	3	37	46.33	252	1346.85	36.70
	Naver	0.32	2.72	4.99	100	72.63	8.52
	Twitter	0	0	0.27	15	1.06	1.03
Malaria	Occurrences	0	0	1.65	14	6.65	2.58
	Naver	5.52	17.24	23.13	100	193.02	13.89
	Twitter	0	3	4.23	34	15.08	3.88
Environ-mental variables	Temperature (°C)	−10.82	14.19	13.27	29.54	94.07	9.70
	Humidity (%)	30.87	67.31	66.70	94.31	172.81	13.15

3. SYSTEM ARCITECTURE

To help foresee whether a patient is experiencing chronic diseaseor not as indicated by his/her medical history. The input esteem isthe attribute value of the patient, which incorporates the patient'sclose to home data, for example, age, sex, the pervasiveness ofside effects, and living propensities (smoking or not) and otherstructured information and unstructured information. The yieldesteem shows whether the patient is experiencing chronic diseaseor not. For disease hazard, demonstrating the precision of riskexpectation relies upon the assorted variety highlight of thedoctor's facility information, i.e., the better is the elementdepiction of the disease, the higher the exactness will be. Forsome straightforward sickness, e.g., hyperlipidaemia, just a coupleof highlights of organized information can get a decent depictionof the illness, bringing about the genuinely great impact of diseaseexpectation. Be that as it may, for an unpredictable disease, forexample, cerebral infarction, diabetes, hypertension, and asthmajust utilizing highlights of structured data isn't a decent method todepict the disease. In this way, use the structured data as well asthe content information of patients given the Support VectorMachine and Naive Bayes (NB) algorithms.In fig. 1, the dataset contains the patient's information related tochronic disease. The dataset is been collected from thehospital. With the help of the dataset, the accurate prediction ofthe disease can be done. In structured data, the prediction ofdisease is done with the help of symptoms of each chronicdisease. The disease prediction is done by the NB algorithm. TheNB algorithm is useful for predicting the probability ofmultiple classes based on various attributes. In this, theprediction of disease is 96% based on the symptoms ofchronic diseases like hypertension, diabetes, cerebralinfarction, and asthma. For Structured data, the system uses atraditional machine learning algorithm, i.e., NB algorithm topredict the disease.NB classification is a simple probabilistic classifier. It requirescalculating the probability of feature attributes. For Structuredinformation, the framework utilizes conventional machine learningcalculation, i.e., NB calculation to anticipate the sickness. NBcharacterization is a straight forward probabilistic classifier. Itrequires figuring the likelihood of highlight properties. A NBclassifier is a simple probabilistic classifier based on applyingBayes' theorem with strong independent assumption. A moredescriptive term for the underlying probability model would bethe self-determining feature model. In basic terms, an NB classifierassumes that the presence of a particular feature of a class isunrelated to the presence of any other feature. The NB classifierperforms reasonably well, even if the underlying assumption isnot true. The advantage of the NB classifier is that it onlyrequires a small amount of training data to estimate the meansand variances of the variables necessary for classification. Inorder to train a Naive Bayes (NB) model for text classification,there is a need to prepare data set. Genetic Algorithm includesprocess of initialization, and then it improves with a repetitiveapplication of mutation, crossover, inversion and selectionoperations. It requires a genetic representation and fitnessfunction. When some user's data is missing then it is beenrecovered by genetic algorithm. In unstructured data, if there ismissing data which is caused by patient's mistake. Then missingdata is been recovered with the genetic algorithm. Theunstructured data mainly focuses on the case study andinterrogation which are given by doctors. The Recurrent NeuralNetwork (RNN) algorithm is used to extract features of the text.The stop words are been removed from the text data and thefeatures are extracted successfully. After text feature extraction,SVM Classier performs classification on the data; it will predictwhether the patient is suffering from chronic disease or not. Withthe help of RNN, unstructured data is been converted intostructured and the prediction of chronic disease is been done. In atraditional neural system it is expected that all inputs (andoutputs) are autonomous of each other. On the off chance thatyou need to foresee the following word in a sentence you betterknow which words preceded it. RNNs are called recurrent on thegrounds that they play out a similar undertaking for eachcomponent of a sequence, with the yield being relied upon thepast calculations. Another approach to consider RNNs is that theyhave a "memory" which catches data about what has been figuredup until now. In principle RNNs can make utilization of data insubjectively long arrangements. The textual features are extractedby RNN. In Fig 1. Xt is the input; Ht is the hidden state which iscalculated based on the previous hidden state and input of thecurrent state. V, W and U are weight matrices, gh is theactivation function, bh is the bias function and Ot is theoutput

$$H_t = g_h(UX^t + VH^{(t-1)} + b_{h)}$$

3. Methods

3.1. Dataset and evaluation methods

In this study, we experimented with dataset collected from the hospital for duration 2013-15 in Wuhan, China. Healthcare data: The unprocessed medical data contains medical images, clinical notes, and 155 personal information of the patients. We noted 20320848 medical records with 31919 patients from 2013 to 2015. This dataset includes patients with chronic diseases like cerebral infarction, hypertension, etc. We extracted total 706 sample text records for train and test the model. Each record contains the patient's detail such as patient's information, patient's readme illness, treatment plan, 160 and doctor records etc. Detail description of the dataset is given by table 1,

Table 1: Initial Statistics of Dataset from grade-A Hospital in Wuhan, China.

Statistics	**2013**	**2014**	**2015**
Number of hospitalized patient	7265	24756	10552
Female	42.88%	50.36%	57.60%
Patients with cerebral infarction	57.12%	49.64%	42.40%
Patients with hypertension	1.47%	1.01%	1.66%
Patients with diabetics	1.06%	1.04%	1.98%

and a sample of clinical note is available by figure 1. As we can see from table1, the proportion of female and male patients have little difference with large number of hospitalized patient in 2014. The ratio of patients in the hospital is also almost same for cerebral infarction, hypertension, and diabetes. But in 2015hospitalized patient numbers of chronic diseases are accounted to be 5.63% of the total hospitalized patient in 2015 which is a little more than in 2013 and 2014, while other diseases have small proportion compared to chronic disease. Therefore, this paper mainly focuses on disease diagnosis for cerebral infarction without considering gender difference because cerebral infarction is the most 170 deadly disease among them.Therefore, finally pre-process clinical notes recorded 706 patients with chronic diseases including cerebral infarction, hypertension, and diabetes and divided patients into two categories only.

Evaluation methods:

Accuracy, f1-measure, recall, and precision are used to measure the performance of the model. Measurement matrix accuracy, f1-measure, recall, and precision will be obtained as follows:

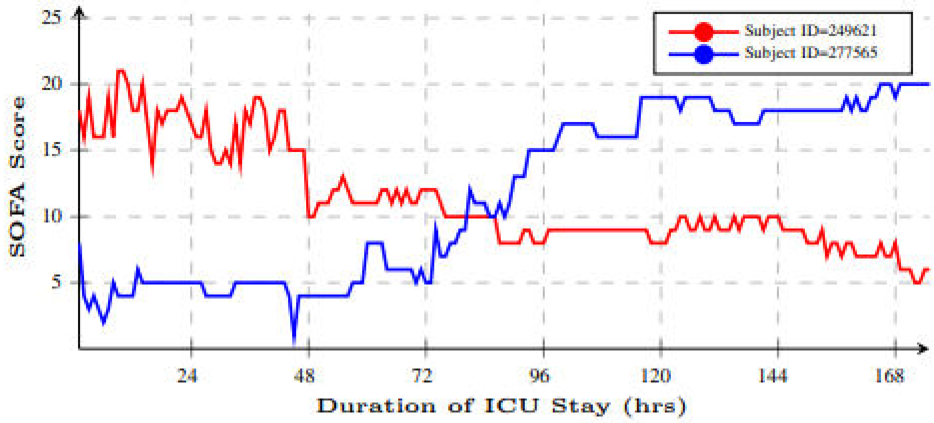
Where true positive are the patients, who are not influenced by the disease and predicted by classifier so. True negative are the patients, who are influenced 180 by the disease and predicted by classifier so. False positive are the patients, who are influenced by the disease and predicted by classifier as not influenced. False negative are the patients, who are not influenced by the disease and predicted by the classifier as influenced.

3.2. Proposed Model

This section describe the model, composed of five components including input layer, embedding layer, self-attention based recurrent convolution, max pooling, full connection, and output layer and uses unstructured medical clinical notes as text data in experiments to predict the cerebral infarction disease risk. Now, we will described each components of model sequentially as follows:

3.2.1. Input Layer

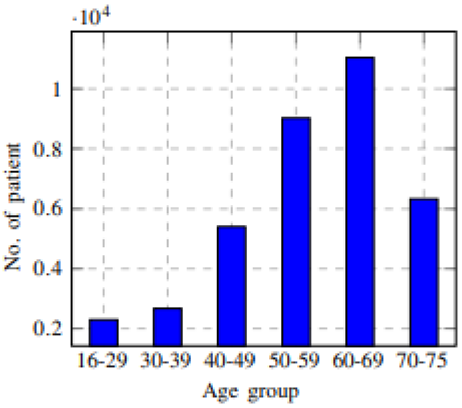
To begin, lets consider after pre-processing we have input text X(x1, x2, ..., xn), XRm∗n, where x1, x2, ..., xn are the n input words, will pass to embedding layer to generate embedding word vectors from corpus text.



4.1 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



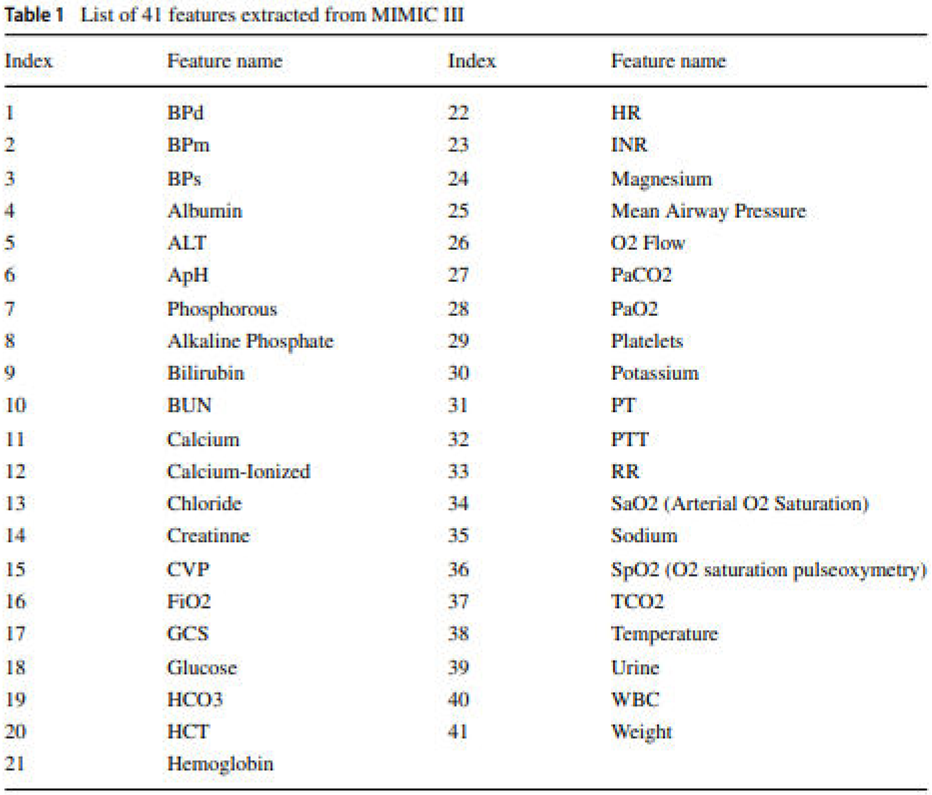


Table 1 List of 41 features extracted from MIMIC III

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 ti to denote the max length in time for the i-th data sample, i = 1, ··· , n. In this way, the data samples can be represented by X = {x1, x2, ··· ,xn}, xi ∈Rti×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 xt,d := xt ,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).

4.2 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 ht 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

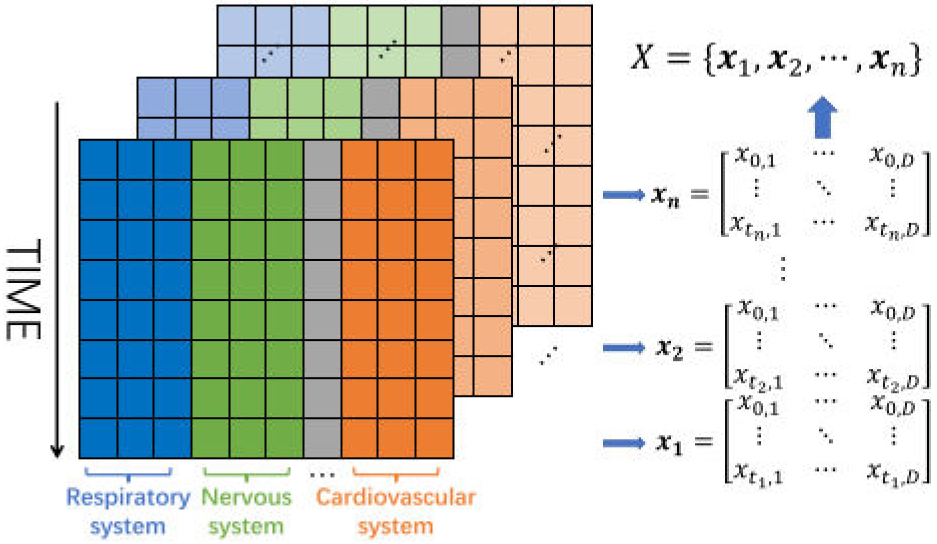

$X = \{x_1, x_2, \cdots, x_n\}$
$x_n = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_n,1} & \cdots & x_{t_n,D} \end{bmatrix}$
$x_2 = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_2,1} & \cdots & x_{t_2,D} \end{bmatrix}$
$x_1 = \begin{bmatrix} x_{0,1} & \cdots & x_{0,D} \\ \vdots & \ddots & \vdots \\ x_{t_1,1} & \cdots & x_{t_1,D} \end{bmatrix}$
TIME
Respiratory system
Nervous system
Cardiovascular system

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:

$$\mathbf{h}_t = \begin{cases} 0 & t = 0 \\ f\ (\mathbf{h}_{t-1}, x_t) & otherwise, \end{cases}$$

where xt is the input of current time-step, and ht−1 is the hidden state of previous time-step.

The LSTM transition equations are defined as follows:

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

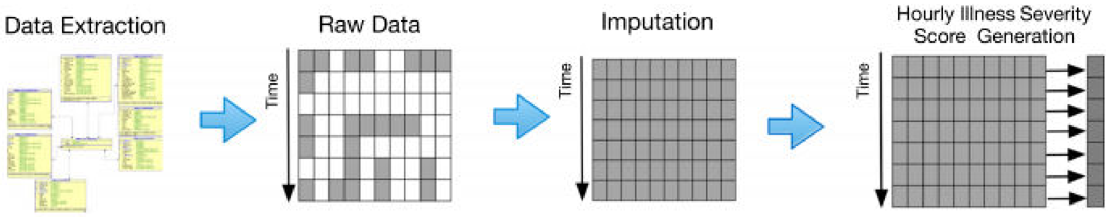


Figure 4 The work-flow of data prepossessing

where xt 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., LST M(m) and LST M(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 LST M(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., LST M(m) and LST M(n) in the figure. The activation function f of the current hidden state for the shared layer, h(s) t , is the same as the one in (2). In contrast, we have modified the activation function for each LSTM h(m) t , which learns different organ features as below:

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

where

denotes an concatenate operation. Meantime, we also change the state c (m) t for in each task-specific LSTM (LST M(m) or LST M(n)) as follows:

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

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 $t-1$.

4.3 Attention mechanism

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

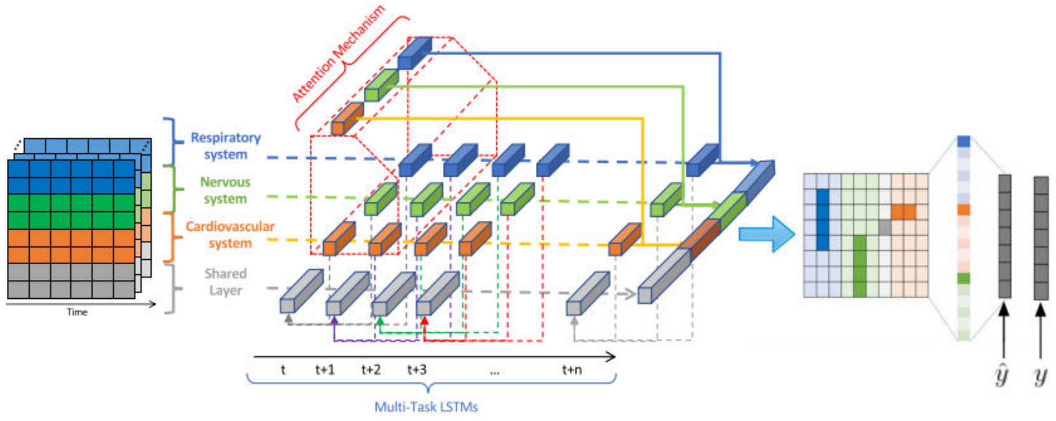


Figure 5 Demonstration of our proposed Multi-task LSTM Architecture with a shared hidden layer. Notethat this structure has not detailed structure of the attention layer

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 corporate 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 ,

, of the hidden state for every feature in the input. Therefore the score for the t-th feature scoret is calculated as follows:

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

where hˆs is the concatenated hidden state of LSTM, in which the t-th the input, and hf is the learned feature of the input. The weight of the t-th input Wt can be computed by using the out put of scoret as follows:

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

here, t means the all input features. The final output af can be computed by using he weight and the hidden state of the feature as a convex sum of hidden states ht :

$$a_f = \sum W_t h_t$$

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:

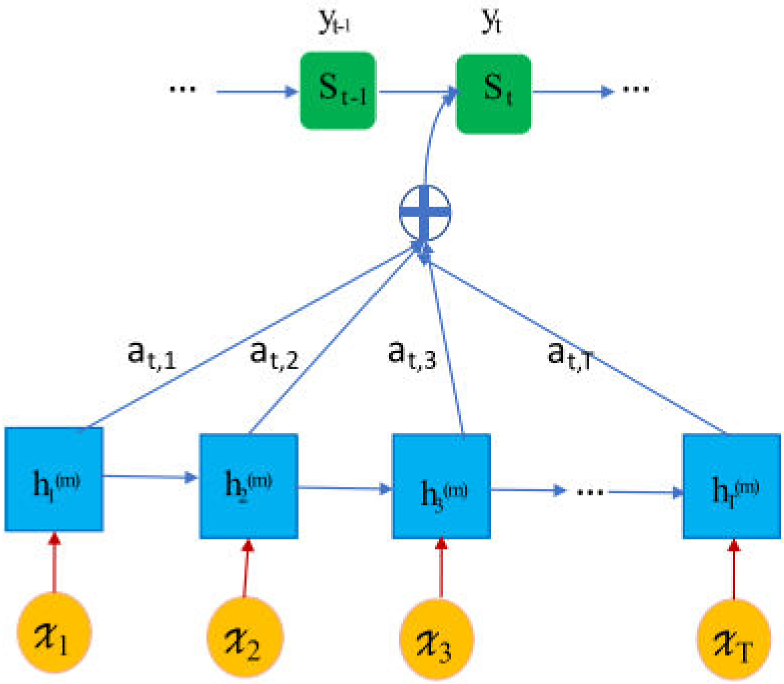


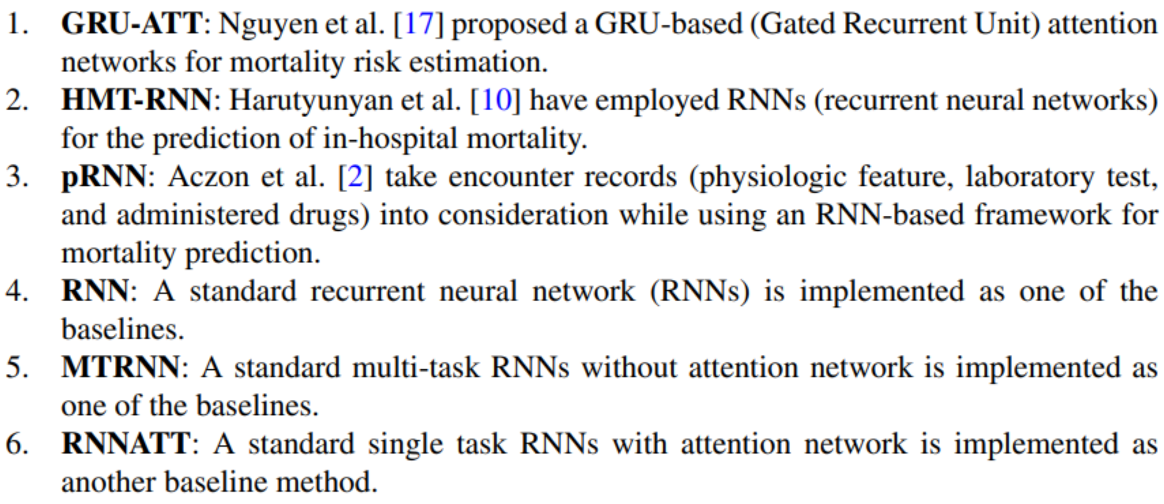
Figure 6 Illustration of the structure of dot-product based attention mechanism

4.4 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.

4.5 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.

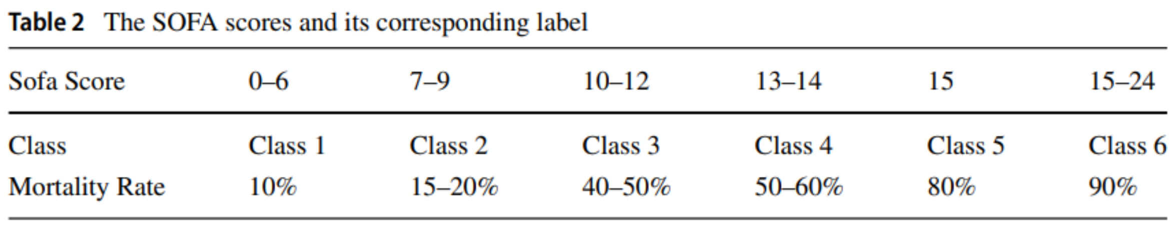


Table 2 The SOFA scores and its corresponding label

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%

5. 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:

$$Accuracy = \frac{\text{TP+TN}}{\text{Total}}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

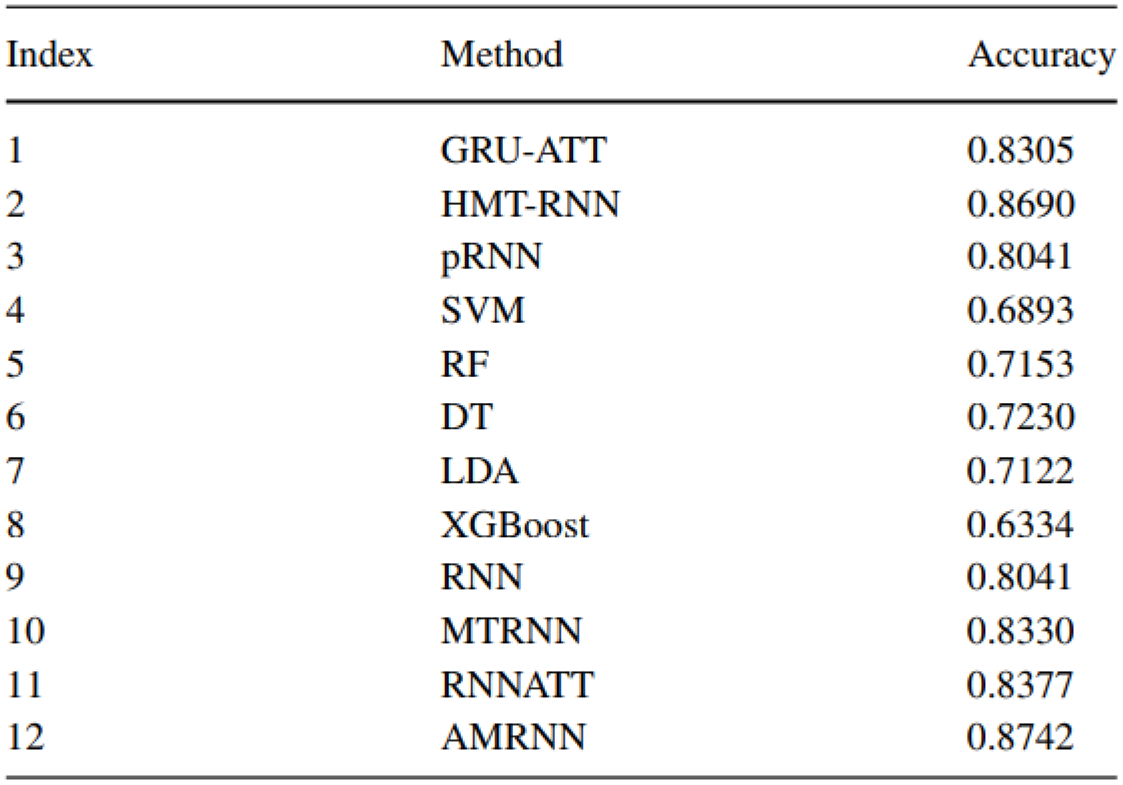
$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

5.1 Discussion

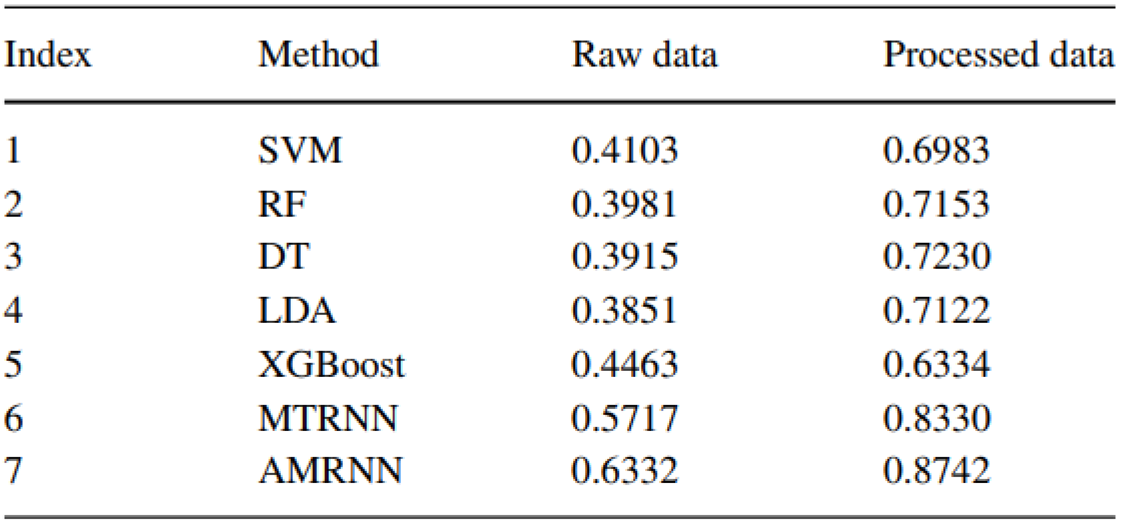
To consider the overall classification performance, we have reported the results of all the methods measured by Accuracy in Table 3. It is clear that our approach performs better than all compared methods. Particularly, the proposed multi-task framework (#12) performs much better than most of the counterparts 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

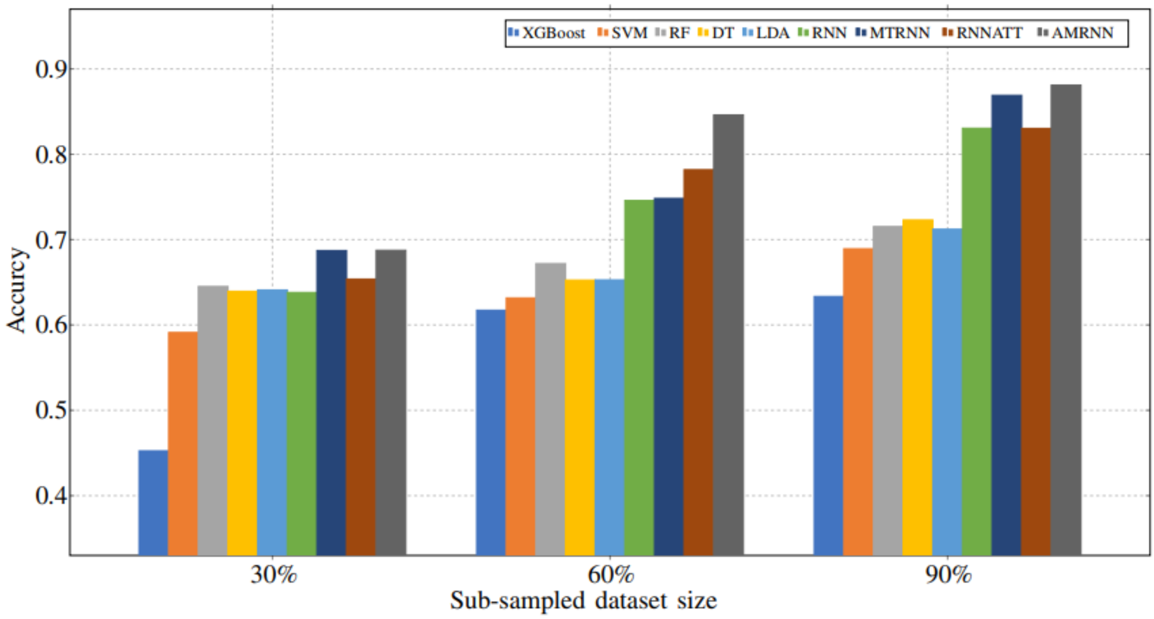


Figure 7 Acc. vs Data Proportion: Prediction accuracy with different training set sizes. x-axis = sub-sampled dataset size; y-axis = accuracy

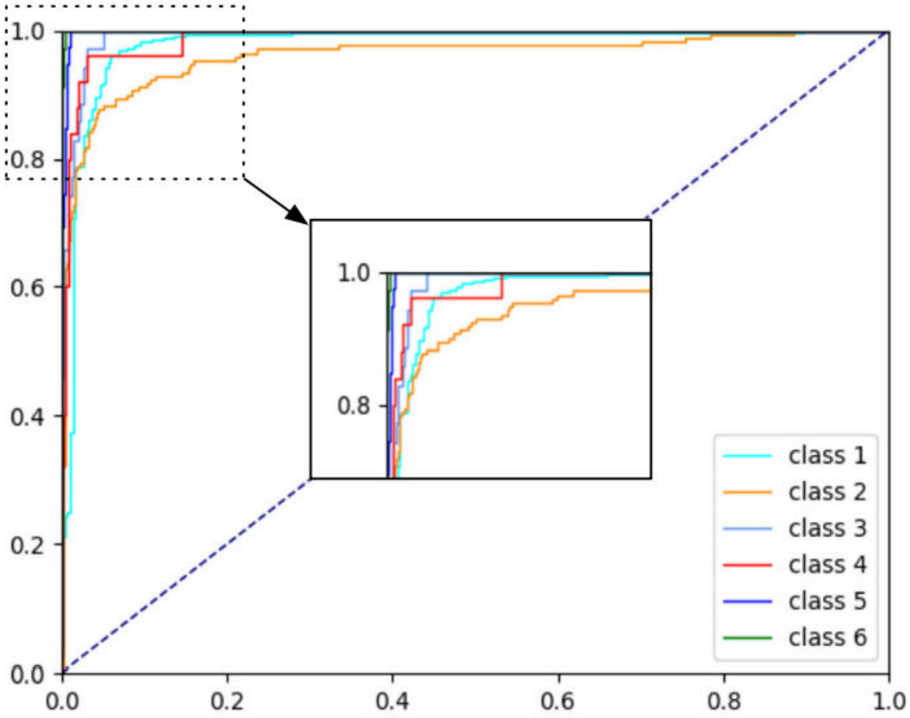


Figure 8 The ROC curves showing the discrimination capability of a classifier. x-axis, True positive size; y-axis, False positive

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

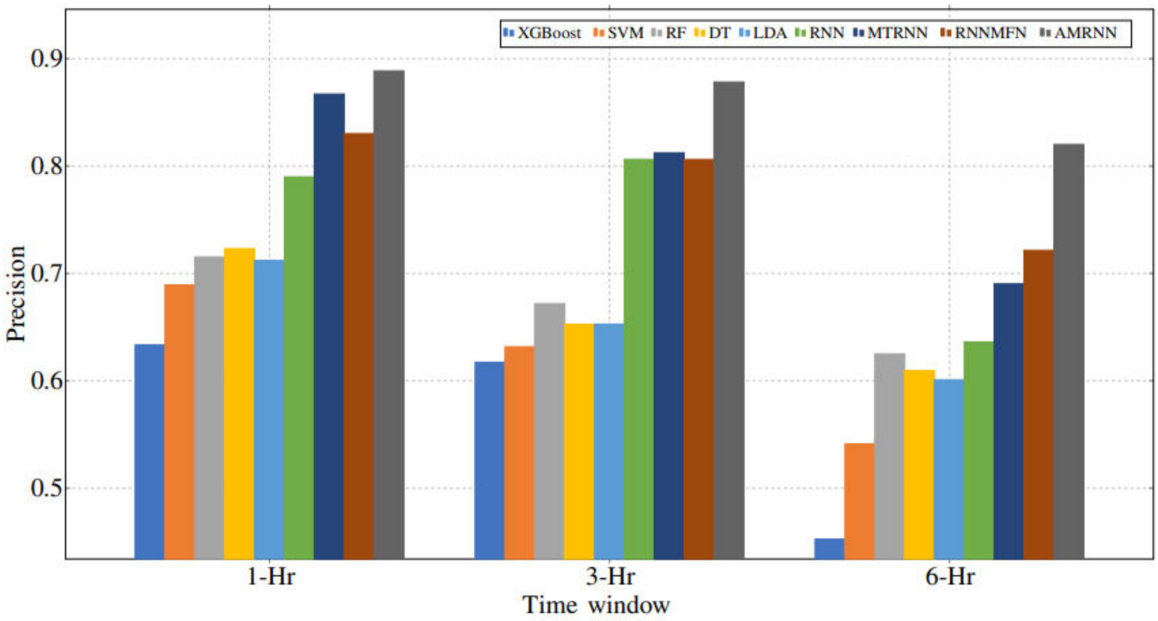


Figure 9 Precision vs Time window: Precision on different length of prediction time windows. x-axis = time window; y-axis = precision

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

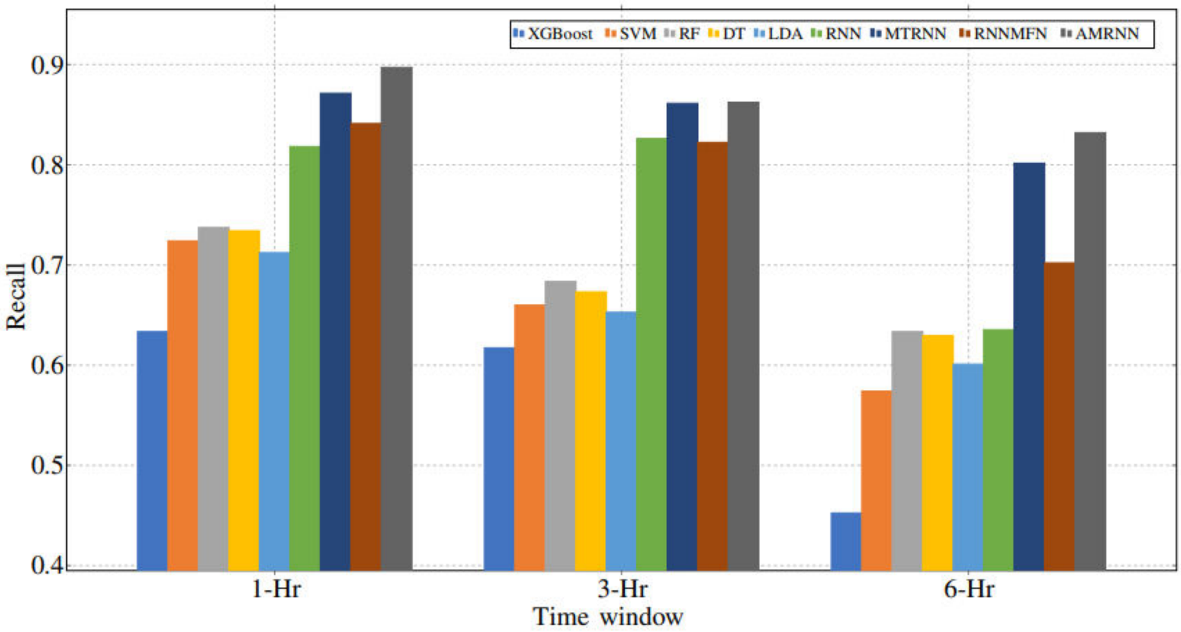


Figure 10 Recall vs Time window: Recall on different length of prediction time windows. x-axis = time window; y-axis = precision

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 compassion methods.

6 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.

7 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. Abdulnabi, 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 (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 (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 americanmulticenter study. Jama 270(24), 2957–2963 (1993)

16. Lipton, Z.C., Kale, D.C., Wetzel, 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., Tackstr ¨ om, 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. Rocktaschel, T., Grefenstette, E., Hermann, K.M., Ko ¨ cisk` ˇ y, T., Blunsom, P.: Reasoning about entailment with neural attention. arXiv:1509.06664 (2015)