





Deep Interpretable Mortality Model for Intensive Care Unit Risk Prediction

Zhenkun Shi^{1,2,3}, Weitong Chen³, Shining Liang^{1,2}, Wanli Zuo^{1,2}, Lin Yue⁴, and Sen Wang³

¹ Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

shizk14@mails.jlu.edu.cn

² College of Computer Science and Technology, Jilin University, Changchun 130012, China

wanli@jlu.edu.cn

³ School of Information Technology and Electrical Engineering, The University of Queensland, Queensland 4072, Australia

⁴ Northeast Normal University, Changchun 130024, China

yuel031@nenu.edu.cn

Abstract. Estimating the mortality of patients plays a fundamental role in an intensive care unit (ICU). Currently, most learning approaches are based on deep learning models. However, these approaches in mortality prediction suffer from two problems: (i) the specificity of causes of death are not considered in the learning process due to the different diseases, and symptoms are mixed-used without diversification and localization; (ii) the learning outcome for the mortality prediction is not self-explainable for the clinicians. In this paper, we propose a Deep Interpretable Mortality Model (DIMM), which employs Multi-Source Embedding, Gated Recurrent Units (GRU), Attention mechanism and Focal Loss techniques to prognosticate mortality prediction. We intensified the mortality prediction by considering the different clinical measures, medical treatments and the heterogeneity of the disease. More importantly, for the first time, in this framework, we use a separate evidence-based interpreter named Highlighter to interpret the prediction model, which makes the prediction understandable and trustworthy to clinicians. We demonstrate that our approach achieves state-of-the-art performance in mortality prediction and can get an interpretable prediction on four different diseases.

Keywords: Data mining · Missing value · Imputation · Deep learning · Healthcare

1 Introduction

Accurate assessment of the severity of a patient's condition plays a fundamental role in acute hospital care especially in ICU, due to the diversity of patients

who mostly suffer from multiple diseases of various types. It is hard to estimate the severity of a patient’s condition in limited time and massive circumstance. However, doctors must give a diagnosis rapidly so that subsequent treatments can be taken.

Traditionally, mortality modeling for ICU patients has been conducted via scoring systems such as the chronic health evaluation (APACHE), sepsis-related organ failure assessment (SOFA) score and simplified acute physiology score (SAPS). All these are adopting fixed clinical decision rules based mainly on physiological data [19]. However, these ICU score systems use only limited measurement indicators to evaluate the mortality risk. For instance, SOFA uses only 11 indicators for scoring. Individually, in MIMIC III [8] there are over 4 thousand indicators. It is obvious that faced with thousands of diseases, these limited indicators are not comprehensive.

Widespread adoption of Electronic Health Record (EHR) and advances in deep learning makes it possible to predict mortality risk more effectively. In fact, several studies using deep learning techniques to forecast in-hospital mortality risk have significantly improved the quality of acute hospital care [20]. However, deep methods are often negotiated as black boxes. While this might not be an obstacle in other more deterministic domains such as image annotation (because the end user objectively validates the tags assigned to the images), in health care, not only the quantity algorithmic performance is necessary but also vital is the reason why the algorithm works. Such model interpretability is crucial for convincing the professionals about actions recommended from the predictive models [11].

By interpreting the mortality risk for the deep model, we mean manifesting textual or visual artifacts that provide the qualitative understanding of the relationship between the clinical measurements, medical treatments, and the model’s prediction. The process of interpreting the mortality risk for a single patient is illustrated in Fig. 1. Cooperation and customization with the clinicians highlighter gives an overall rating that indicates the patient’s mortality risk and integrates, illustrates, and highlights the related events, items, as well as the diagnosis. With these explanations, clinicians can take further actions based on the results and the explanations. Furthermore, it has been observed, for example, that providing explanations can increase the acceptance of movie recommendations [7].

In this paper, we proposed a Deep Interpretable Mortality Model (DIMM) to predict and interpret the ICU patients’ mortality risk. We evaluate the DIMM on MIMIC-III and show that it is highly competitive and outperforms the state-of-the-art traditional methods and commonly used deep learning methods. Furthermore, we can provide evidence for our prediction model in convincing the clinicians to trust the prediction Here is a summary of our contribution: (1) **Multiple Perspectives for ICU Mortality Risk Formulation.** We formulate ICU mortality prediction as a multi-source and multi-task learning problem, where sources correspond to clinical measurements and medical treatment, tasks correspond to diseases. Our model enables us to incorporate disease-specific con-

text into mortality modeling. (2) **Use more inclusive data set.** Use of the entire data set of ICU patients without filtering on length-of-stay. Other similar submission, and previous works heavily filter the data set. (3) **Explainability of the Prediction Outcomes.** We add a highlighter to provide evidence-based trustworthy interpretation to the deep model, and this is very important in real life situations. (4) **Comprehensive Evaluated Experiments.** We demonstrate the effectiveness of our method by using MMIC-III benchmark dataset and achieve the state-of-the-art performance.

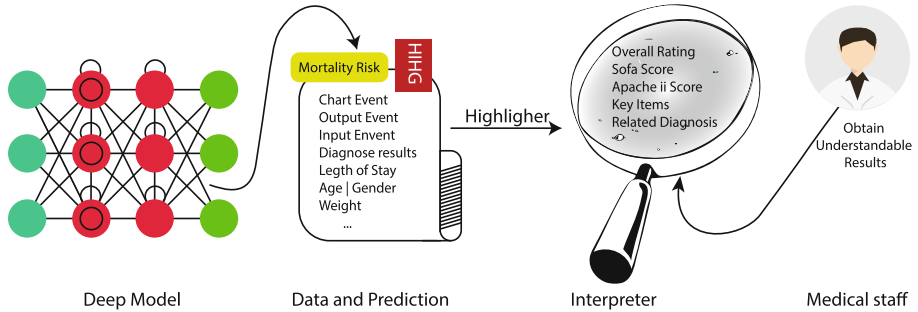


Fig. 1. The process of interpretation.

2 Related Work

Mortality risk prediction has a long history in medical domain, traditional methods for mortality modeling mostly based on scoring systems. However, it is pointed out that these ICU scoring systems are used for only a few patients, namely, 10%–15% of US ICU patients as of 2012 [4]. As information of a more varied, time-evolving nature became available as a form of EHR such as clinical notes and intervention records, data-driven predictive modeling has been explored extensively in recent year [3, 6, 12, 18]. However, these data mining and statistical learning approaches typically need first to perform feature engineering to obtain adequate and more robust features from those data and then build prediction or clustering models on top of them [5, 23]. More recently, RNNs provide new effective paradigms to obtain end-to-end learning models from complex data. Harutyunyan [6] and Song [20] used LSTM and Attention Model, respectively, to predict in-hospital mortality and provides the state-of-the-art performance. However, a large body of current works [3, 6, 20] for mortality prediction mainly focuses on improving the ability of classifiers. Moreover, scarcely any publications think more about the reality in an application scenario, beyond the prediction ability and accuracy; the clinician is caring much about how the model predicted so that they can choose to trust the prediction results or not. What's more, the clinician needs to know the irregular figures and the possible treatment measures. Therefore, low interpretability is a common problem of deep neural networks,

but all the works mentioned above have not tackled this problem. Ribeiro *et al.* [14] did some significant work in explaining the prediction, but his work is not evidence-based and is not fully effective in that medical field. Lipton [10] points out that model interpretability in machine learning has multiple definitions but none can be appropriately applied in mortality prediction. Ahmad *et al.* [1] give a comprehensive summary of model interpretability in machine learning. Combining the term of Evidence-based Medicine and previous works, in this paper, the interpretation given evidence and intuitive explanations on how we made the prediction.

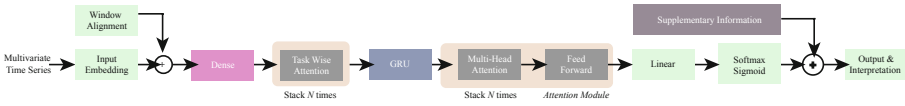


Fig. 2. An overview of the proposed GRU-based framework for mortality risk prediction and interpretation.

3 Proposed Approach

In this section, we describe our framework, DIMM.

3.1 Problem Statement

For a given ICU stay of length T hours, it is assumed that we are given a series of regularly clinical actions $\{x_t\}_{t>1}^T$, where x_t is a vector of clinical action at time t . $x_t = a_{ti}\Theta b_{tj}$, where a_{ti} stand for clinical observation vector at time t_i and b_{tj} represent the clinical treatment vector at time t_j , Θ is a joint operation of vector a_{tj} and b_{tj} . Our objective is to generate a sequence-level prediction and give an interpretation of each prediction. The type of prediction depends on the specific task and can be donated as a discrete scalar vector Y for multi-task classification, each discrete vector Y_i for the regression problem. The proposed framework is shown in Fig. 2.

3.2 Input Embedding

Give the R actions for each step t , the first step is to generate an embedding that captures the dependencies across different diseases without the temporal information. Here, N denote the number of diseases. The mortality prediction model is constructed for each disease. The n -th disease has P_n patients, and p -th patients with n -th disease is associated with two feature vectors A_p^n and B_p^n derived from the EHR, where A_p^n donates the clinical measurements and B_p^n donates the medical treatments. The dimension of A and B are α and β , respectively. Combining A_p^n and B_p^n , we generated a new feature vector Φ^n for the n -th disease:

$$\Phi^n \equiv [\phi_1^{(n)}, \phi_2^{(n)}, \dots, \phi_{P_n}^{(n)}]^T \tag{1}$$

$$\phi_p^n = \lambda_1 A_p^n \Theta \lambda_2 B_p^n \quad (2)$$

where Θ is the linear combine operation.

In order to interpret our prediction in the later process, we added a location mask LM to accompany the input embedding process. Because with the process of a series map, transformation and dropout, it is hard to map the result to the original vector space. We define $LM^n = f(\phi_p^n)$, f is the mapping function:

$$f(\phi_p^n) = \begin{cases} 0, & \text{if } \phi_p^n = 0 \\ p, & \text{otherwise.} \end{cases} \quad (3)$$

3.3 Window Alignment

Since our framework contains multiple actions, medical treatments, and clinical measurements. The medical treatments will take a while to take effect and influence the measurement results. Assume $A_p^n \circ t_i$ represent the clinical measurement at time step t_i and $B_p^n \circ t_j$ represents the medical treatment at time step t_j . The alignment is performed by mapping $P(\phi)$ and $P(\phi)$ into a unique time step $P(\phi)$. This strategy is fully competent in our tasks. Besides, in the same time window $P(\phi)$, t_j is usually later than t_i , and this accords with the prevailing medical sense.

3.4 Dense Layer

To balance the computational cost as well as the predictable performance, we need to reduce the dimensions before we transfer the raw medical data to the next process step. The typical way is simply to concatenate an embedding at every step in the sequence. However, due to the high-dimensional nature of the clinical features, this causes ‘‘cursed’’ representation which is not suitable for learning and inference. Inspired by the Trask’s work [21] in Natural Language Processing (NLP) and Song’s [20] in clinical data processing, we add a dense layer to unify and flatten the input features as well as keep the interpretability. To prevent overfitting, we set dropout = 0.35 here.

3.5 Task Wise Attention

Inspired by BiDAF [16] in NLP field, after the dense layer, we add a Task wise attention layer to linking and fusing information from the medical treatments and the clinical measurements. The inputs to the layer are treatments vector $P(\phi)$ and measurements vector $P(\phi)$, these two vectors have been flattened in the previous layer. The output clinical action vector $\tilde{\phi}_p^n$ along with the input embeddings from the previous layer. In this layer, we compute attentions in 4 directions: clinical measurements self-attention (M2M); (1) medical treatment self-attention (T2T); (2) measurements to treatments (M2T); (3) treatments to measurements (T2M).

The first two self-attentions can help us focus on the most critical part of their self-vector spaces. For example, cardiotoxic can be very helpful in some emergency case, and this treatment is crucial in the whole ICU process. However, this is a discontinuous treatment and usually appears once in the time series; by using self-attention mechanism we are able to capture this vital procedure and raise its weight in the prediction. The M2T and T2M attention are derived from shared similarity matrix $S \in \mathbb{R}^{A \times B}$ between the input embedding of clinical measurements (A) and medical treatment (B), where S_{ab} indicates the similarity between a -th clinical measurement and b -th medical treatment. The similarity matrix is computed by

$$S_{ab} = \ell(A_{:,a}, B_{:,b}) \quad (4)$$

Where ℓ is a trainable scalar function that encodes the similarity between two input vector, and $A_{:,a}$ is the a -th column vector of A, and $B_{:,b}$ is the b -th column vector of B. We choose $\ell(A, B) = w_{(S)}^T [A; B; A \circ B]$, where $w_{(S)}^T$ is a trainable weight vector, \circ is elementwise multiplication, $[;]$ is vector concatenation across row, and implicit multiplication is matrix multiplication. By introducing S we can obtain the attention and the attended vectors in both directions.

In addition, this attention mechanism in our work is task wise. We add a tunable switch $S_{w,i}$:

$$S_{w,i} = \begin{cases} 0, & \text{if can improve the prediction performance} \\ 1, & \text{otherwise.} \end{cases} \quad (5)$$

where $w \in \{M2M, T2T, M2T, T2M\}$ and i refers the predict task. The motivation why we adopt the task wise mechanism is that the measurements and treatments from different disease categories may hugely different. For instance, the diagnosis and the treatments of the diseases of the respiratory system and the diseases of the genitourinary system are nearly different. So it is hard to share feature spaces between these two kinds of conditions. Therefore, by adopting this mechanism, we not only can reduce the training difficulty but also can avoid inducing noises.

3.6 The Gated Recurrent Unit Layer

The GRU takes the sequence of action $\{x_t\}_{t \geq 1}^T$ from the previous dense layer and then associate p -th patient with a binary class label $y_{n,p}$, denotes the class label for the p -th patient with the n -th disease. $y_{n,p}$ is set as follows:

$$y_{n,p} = \begin{cases} 0, & \text{if dead within 60 days after ICU} \\ 1, & \text{otherwise.} \end{cases} \quad (6)$$

We create a P_n -dimensional response vector for the n -th disease:

$$y^{(n)} = (y_{n,1}, y_{n,2}, \dots, y_{n,P_n})^T \quad (7)$$

For the ICU patients’ mortality risk prediction, we adopted GRU and represent the posterior probability of the outcome of patient p being death as:

$$\Pr[y_{n,p} = 0 | \phi_p^{(n)}] = \sigma(\omega^{(n)N} \phi_p^{(n)}) \tag{8}$$

where $\sigma(a)$ is the sigmoid function $\sigma(a) \equiv (1 + \exp(-a))^{-1}$ and $\omega^{(n)}$ is a $\alpha + \beta$ dimensional model parameter vector for the n -th disease.

To learn the mutual information of data resulting from the customization, we learn models for all diseases jointly, so that we can share the same segment information across the diseases. We represent the trainable parameters of the GRU as a $(\alpha + \beta) \times N$ matrix $W \equiv [\omega^{(1)}, \omega^2, \dots, \omega^n]$.

3.7 Multi-head Attention and Feedforward

This attention layer is designed to capture dependencies of the whole sequence. In the ICU scenario, the actions closer to the current position are more critical than the farther one. And we should consider information only from positions earlier than the current position being analyzed. Inspired by [22], we use H-heads attention to create multiple attention graphs, and the resulting weighted representations are concatenated and linearly projected to obtain the final representation. Moreover, we also add 1D convolutional sub-layers with kernel size 2. Internally, we use two of these 1D convolutional sub-layers with ReLU (rectified linear unit) activation in between. Residue connections are used in these sub-layers. Unlike [20] and [6] making mortality predictions only once after a specific timestamp, we give prediction and interpretation at each timestamp. This is more helpful for the ICU clinicians because they need to know the patients’ mortality risk at any time other than at the particular time. We stack the attention module N times and use the final representations in the mortality risk prediction model.

3.8 Linear and Softmax Layers

The linear layer is designed to obtain the logits from the unified output of attention layer. The activation function used in this layer is ReLU. The last layer is preparing for the output based on different tasks. We use sigmoid for the binary mortality task, the loss function is:

$$Loss_m = -(y \log(\bar{y})) + (1 - y) \cdot \log(1 - \bar{y}) \tag{9}$$

where y and \bar{y} denote the true and predicted labels.

We use softmax to distinguish between N different diseases, and the loss function is:

$$Loss_d = \frac{1}{N} \sum_{n=1}^N -(y_k \cdot \log(\bar{y}_k) + (1 - y_k)). \tag{10}$$

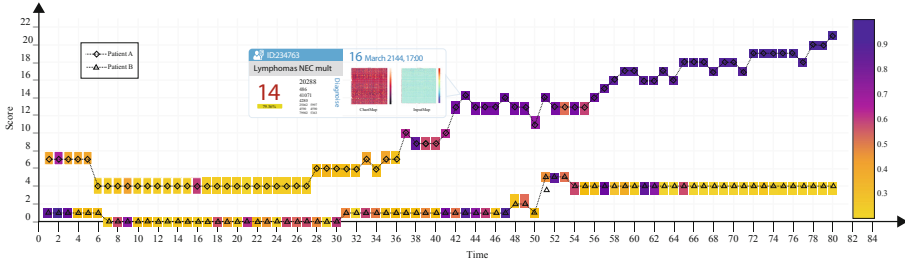


Fig. 3. Deep interpretable mortality prediction framework prediction results showcase. (Color figure online)

3.9 Focal Loss

Due to the distribution otherness of different mortality is very large, and this led to an extreme sample imbalance problem. For example, in MIMIC III dataset “Respiratory distress syn” ICU mortality risk is 3.06% and “Angina pectoris NEC/NOS” ICU mortality risk is 23.2%. This problem is very similar to the sample imbalance problem in text summarization [9]. Therefore, the training difficulty of each disease is different. Moreover, the training difficulty of each action in different diseases is also not the same, represented by $P(\phi)$. The $P(\phi)$ means that an action is an easy training disease where $P(\phi)$ is closed to 1 because the model can predict it with the confidence of one hundred percent, while it will be a difficulty training disease if $P(\phi)$ is small to zero.

Consequently, we need to assign a soft weight to each disease’s loss for the model and pay more attention on those whose $P(\phi)$ is small for getting better performance on it. Inspired by the work in NLP [17], we improved our loss function by introducing focal loss. Like Lin’s work [9], we do use cross-entropy (CE) to define $P(\phi)$. Let p_f denote the model’s estimated probability for the class with the label $y_{n,p} = 0$, and in this work means p -th patient with n -th disease and eventually death. We define $P(\phi)$ as:

$$P(\phi) = \begin{cases} -\log(p_f), & \text{if } y_{n,p} = 0 \\ -\log(1 - p_f), & \text{otherwise.} \end{cases} \quad (11)$$

Then we add a modulating factor $(1 - P(\phi))^\gamma$ to the entropy loss, with tunable focusing parameter $\gamma \geq 0$. We define the focal loss as:

$$FL(P(\phi)) = -\lambda_\phi(1 - P(\phi))^\gamma \log(P(\phi)) \quad (12)$$

where λ_ϕ and γ are hyperparameters, which are set as 0.25 and 1 respectively in our experiment.

3.10 Output and Interpretation

Incorporated with the supplementary information from original source, this layer is designed to generate the understandable prediction results. We give the pre-

dictions at each time step and the time span can be controlled by the end user. Therefore, the clinician could get the updated patients’ mortality risk at any time.

Figure 3 is a showcase of the DIMM prediction with evidence interpretation results. The x-axis is the ICU stay time sequence and the y-axis is the patient’s ICU Scoring System score. Here, we plot the SOFA score. *Patient B* is the current patient and *Patient A* represent the similar patient who is like *B*. The color donates the mortality risk of the patient at current time slot, the darker color indicates the higher mortality risk. If we focus on a specific time slot, the Highlighter will give detailed information and the evidence of how we got this prediction. The outputs includes the necessary information about the current patient, the diagnosis, and diagnostic order, the mortality risk, the severity scores from different score systems, the heat map of crucial clinical measurements and, the heat map of adequate medical treatment. Moreover, we also give the most similar patients compared to the current one, and this is very helpful to the clinician.

4 Experiment

4.1 Dataset

We use real-world datasets from MIMIC-III to evaluate the proposed approach. And, we treat each ICU stay as a single case. In other words, different ICU stays of the same patient will be treated as separate cases, and this will help us to get more samples. The diseases choosing standard is like Nori [13]. According to the International Classification of Diseases (ICD) codes, we extracted the following diseases for each patient: the primary diseases that caused the patients admission, and comorbidities the patient had at the time of admission, where the number of comorbidities is at most ten. After filtering the patients aged below 16, we obtained 30508 patients for this study. The total diagnosed diseases in these patients id 5395. For the features, we included 1529 clinical measurement features and 330 medical treatment features. In this study, we picked out the most common four diseases as our prediction tasks, 4019 (Essential hypertension), 41401 (Coronary atherosclerosis), 25000 (Diabetes mellitus), 5849 (Acute renal failure), the sample size of each task is 15561, 9716, 6480, 6270, respectively.

4.2 Prediction Settings

We adopted 30 days mortality as a time window to measure the mortality. That is, if a patient died within 30 days after his or her ICU stay, the outcome is “death” and otherwise “survival”. As a measure of different diseases, if the prediction results match one of the main caused diseases, the outcome is “true” otherwise “false”. We predict every step. The learning rate we use is 0.001 and epochs size: 30. In our experiment batch size: 32, ADAM dropout to 0.35 and learning rate at 0.001. In the single task process, we set the integrated attention stack time $N = 4$. In the multi-head layer head = 4, and, the in the task of 5849, 2500 is $N = 4$, in the task of 41401, 4019 is $N = 2$. In the multitask process, we set $N = 4$.

Table 1. Disease-specific mortality risk prediction tasks

Task	Positive	Negative	Train	Validation	Test
4019	13321	2240	11507	2334	1719
41401	8667	1048	7184	1457	1074
25000	5295	1185	4792	972	716
5849	4433	1837	4637	941	693

4.3 Compared Methods

We compared our proposed method with the following seven methods. Four traditional methods, Logistic Regression with L2 regularization, Random Forest and XGBoost, SVM. Three NN-based methods, *Temporal Convolutional Networks* (TCN) [2], MIMIC-III benchmark tasks (BM) [6], Attend and Diagnosis (SAnD) [20]. To ensure all methods use the same data, we fixed the training and testing dataset. The validation and test data we use is approximately 30% of the whole dataset. The detailed information is shown in Table 1.

4.4 Evaluation Metric

As listed in Table 1, all tasks are facing data unbalancing problem. To comprehensively assess our DIMM framework, we use four different evaluation metrics in our experiment. First, we used Area under Receiver Operator Curve (AUROC) to evaluate our model, which is a combined measure of sensitivity and specificity. The next primary metrics for evaluation is Area under Precision-Recall Curve (AUPRC) because the precision-recall plot is more informative than the Receiver Operating Characteristics (ROC) plot when evaluating binary classifiers on imbalanced datasets [15]. We also considered the Accuracy (ACC).

4.5 Ablation Study

To demonstrate the synergy between different layer modules for DIMM architecture, we trained the different sub-modules of DIMM separately and conducted ablation comparison. The experiment results are shown in Table 2. From the table, we can find that the full DIMM framework can obtain the best result at most of the time. From the columns, we can conclude that the window alignment, the dense layer and the GRU layer composed the backbone of the DIMM. In these four prediction tasks, if we remove some attention layer during the training process, some evaluation metrics may perform better suggesting that not all kinds of attentions are working for many reasons such as different diseases from different categories may prove different in measurements and treatments or history of a patient’s diseases is uncorrelated to the current ones. So the task wise attention mechanism is essential. Due to limited space we show only AUROC and ACC here.

Table 2. Ablation study results of different layers

Task	Metric	No-wA	No-dense	No-interact	No-internal	No-TWA	No-GRU	No-STA	FULL
4019	AUROC	0.9009	0.9118	0.9285	0.9284	0.9328	0.8094	0.9329	0.9329
	ACC	0.8965	0.9086	0.9333	0.9295	0.9299	0.8559	0.9789	0.9789
41401	AUROC	0.9403	0.9237	0.9487	0.9450	0.9427	0.8406	0.9483	0.9473
	ACC	0.9450	0.9264	0.9463	0.9428	0.9386	0.8727	0.9821	0.9821
25000	AUROC	0.8955	0.9077	0.9354	0.9446	0.9321	0.8044	0.8933	0.9449
	ACC	0.8952	0.8785	0.9271	0.9281	0.8893	0.8088	0.9502	0.9502
5849	AUROC	0.8934	0.9086	0.9421	0.9403	0.9330	0.8054	0.9247	0.9421
	ACC	0.8326	0.8538	0.9017	0.8953	0.9117	0.7686	0.9763	0.9763

No-wA: eliminate the window alignment. No-dense: eliminate the dense layer. No-interact: eliminate the M2T or T2M attention. No-internal: eliminate the M2M or T2T attention. No-TWA: eliminate the whole task wise attention layer. No-GRU: eliminate the GRU layer. No-STA: eliminate the attention layer after GRU layer. FULL: full DIMM framework.

4.6 Results and Discussion

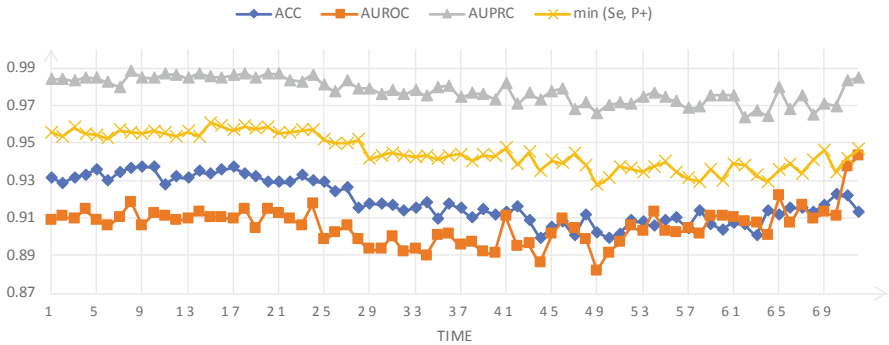
**Fig. 4.** The performance of the mortality prediction sequence.

Table 3 shows the prediction results. We can see that our model significantly outperformed all the baseline methods. We conducted our experiment for the 30 days mortality risk prediction. To simulate real conditions in ICU, we use all ICU stays. We include all the ICU patients without considering the length of their stay. This is unlike the previous work Harutyunyan et al. [6] and Song et al. [20], who are using measurements from the last 24 h. The performance of different evaluation indicators are as shown in Fig. 4. It is clear that our framework is very stable among all the evaluation metrics.

We can see that NN-based methods are outperforming much better than the LR, SVM, and RF. This suggests that the neural network is much powerful than the traditional methods in our tasks, as we expected. We use the same model and the same settings on different tasks and the results are notably different, suggesting that there exists a notable diversity of the diseases, as we mentioned in

Table 3. Results for mortality risk prediction task

Metrics	Methods									
	LR	RF	XGBoost	SVM	TCN	SAnD	BM-s	BM-m	DIMM-s	DIMM-m
Task1: 4019 Essential Hypertension										
AUROC	0.7601	0.8110	0.8601	0.7702	0.8129	0.8436	0.8906	0.8948	0.9186	0.9369
AUPRC	0.9391	0.9546	0.9659	0.9386	0.9518	0.9658	0.9723	0.9727	0.9774	0.9807
ACC	0.8402	0.8546	0.8637	0.8406	0.8435	0.8833	0.8888	0.8996	0.9219	0.9789
Task2: 41401 Coronary Atherosclerosis										
AUROC	0.8144	0.8313	0.8775	0.7990	0.8356	0.7950	0.9066	0.9088	0.9267	0.9473
AUPRC	0.9513	0.9619	0.9676	0.9469	0.9584	0.9635	0.9749	0.9775	0.9806	0.9854
ACC	0.8548	0.8691	0.8818	0.8496	0.8678	0.9000	0.9009	0.9015	0.9291	0.9821
Task3: 25000 Diabetes Mellitus										
AUROC	0.7414	0.7798	0.8475	0.7907	0.7749	0.8234	0.8581	0.8954	0.8785	0.9499
AUPRC	0.8973	0.9217	0.9420	0.9181	0.9094	0.9465	0.9466	0.9616	0.9540	0.9748
ACC	0.7796	0.8002	0.8319	0.7820	0.7963	0.8335	0.8535	0.8863	0.8657	0.9502
Task4: 5849 Acute Renal Failure										
AUROC	0.8000	0.7900	0.8368	0.7673	0.7201	0.7844	0.8535	0.8639	0.8933	0.9421
AUPRC	0.8671	0.8699	0.8989	0.8356	0.8009	0.8811	0.9070	0.9079	0.9235	0.9741
ACC	0.7714	0.7563	0.7865	0.7191	0.7091	0.7992	0.8132	0.8258	0.8644	0.9763

 Non-NN based
 NN based
 Ours
 SAnD: attend and diagnosis, BM: benchmark, -s:single-task, -m: multi-task

the introduction. The fact that all multi-task models can get better performance than the single task models indicates that joint inferencing with multiple related tasks can lead to superior performance in each of the individual tasks, while drastically improving the training. Hence, it is essential to building the mortality prediction model in a multi-task way.

That our single task DIMM outperformed the multi-BM indicates that, diseases specific assessment is helpful in improving the prediction performance. Thus, we can generalize our model to predict ICU mortality risk according to different diseases. More importantly, detailed disease specific predictions are more significant than the disease nonspecific ones, because respiratory physicians are not dealing with the otolaryngology patients and the domain knowledge between respiratory and otolaryngology is different.

In the interpolation process, as shown in Fig. 5, first we show how we made the prediction. The color represents the contribution of current actions; the further from 0, the more significant the contribution is. We can infer from that for the given patient and a specific disease, there are always some critical clinical measurements and medical treatment. Besides evidence-based interpretation figures, we also manifest the textual and visual artifacts to provide the qualitative understanding of the relationship between the clinical measurements, medical treatments, and risk prediction. As shown in Fig. 3: DIMM framework prediction results, this is a showcase that can transfer the information clearly and effectively to the clinicians. So the clinicians can make explicit and judicious

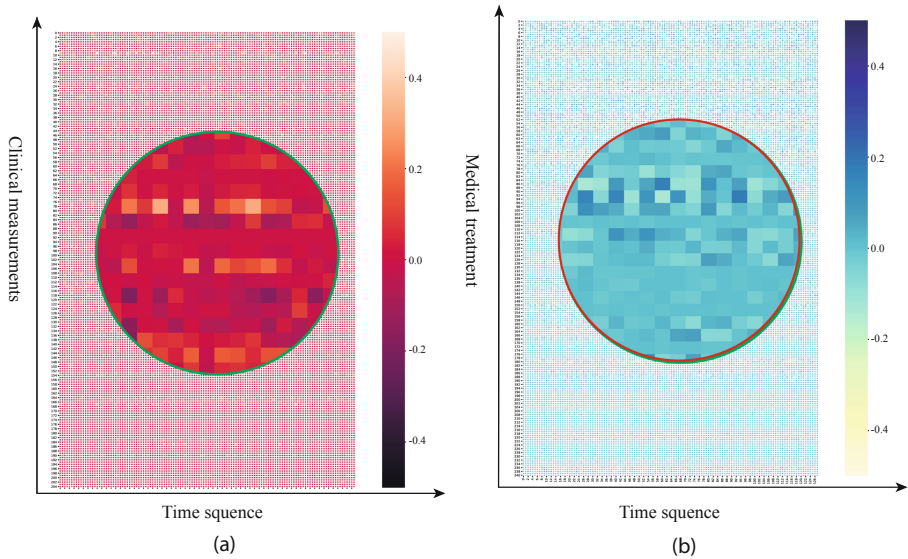


Fig. 5. The evidence-based interpretation of clinical measurements (a) medical treatment (b); mortality prediction. The color represents contribution of the current measurement. (Color figure online)

use of current best evidence in making decisions about the care of individual patients.

5 Conclusion

In this study, we presented a new ICU mortality prediction model, DIMM. The significances of our proposed model can be identified as: (1) We considered the diversity of diseases. This accords with the medical situations. (2) By introducing embedding we can utilize multi-sources for prediction and this has two advantages: improving performance and doing interpretation in the later process. (3) Two attention layers help us capture both the internal correlation between the measurements and the treatments. (4) By using focal loss function, we alleviated problem caused by the unbalanced dataset in the training process. (5) Explainability for the deep model is realized by using the Highlighter. All clinical decisions are based on evidence, we provide a visualization view to the clinicians for our model. This is crucial to the clinical cohort because further medical actions are based on trust chains of the whole prediction process other than a single digit. Nevertheless, how to evaluate the interpretation still remains a challenge in scientific research, and future work can focus on this problem.

References

1. Ahmad, M.A., Eckert, C., McKelvey, G., Zolfaghar, K., Zahid, A., Teredesai, A.: Death vs. data science: predicting end of life. In: AAAI (2018)
2. Bai, S., Kolter, J.Z., Koltun, V.: Convolutional sequence modeling revisited (2018)
3. Bhattacharya, S., Rajan, V., Shrivastava, H.: ICU mortality prediction: a classification algorithm for imbalanced datasets. In: AAAI, pp. 1288–1294 (2017)
4. Breslow, M.J., Badawi, O.: Severity scoring in the critically ill: part 1—interpretation and accuracy of outcome prediction scoring systems. *Chest* **141**(1), 245–252 (2012)
5. Gil, V., et al.: Emergency heart failure mortality risk grade score performance for 7-day mortality prediction in patients with heart failure attended at the emergency department: validation in a spanish cohort. *Eur. J. Emerg. Med.* **25**(3), 169–177 (2018)
6. Harutyunyan, H., Khachatryan, H., Kale, D.C., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. arXiv preprint [arXiv:1703.07771](https://arxiv.org/abs/1703.07771) (2017)
7. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
8. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016)
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018)
10. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint [arXiv:1606.03490](https://arxiv.org/abs/1606.03490) (2016)
11. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **19**, 1236–1246 (2017)
12. Nori, N., Kashima, H., Yamashita, K., Ikai, H., Imanaka, Y.: Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864. ACM (2015)
13. Nori, N., Kashima, H., Yamashita, K., Kunisawa, S., Imanaka, Y.: Learning implicit tasks for patient-specific risk modeling in ICU. In: AAAI, pp. 1481–1487 (2017)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
15. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**(3), e0118432 (2015)
16. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint [arXiv:1611.01603](https://arxiv.org/abs/1611.01603) (2016)
17. Shi, Y., Meng, J., Wang, J., Lin, H., Li, Y.: A Normalized Encoder-Decoder Model for Abstractive Summarization Using Focal Loss. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2018. LNCS (LNAI), vol. 11109, pp. 383–392. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99501-4_34
18. Shi, Z., Zuo, W., Chen, W., Yue, L., Hao, Y., Liang, S.: DMMAM: deep multi-source multi-task attention model for intensive care unit diagnosis. In: Li, G., Yang, J., Gama, J., Natwichai, J., Tong, Y. (eds.) DASFAA 2019. LNCS, vol. 11447, pp. 53–69. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18579-4_4

19. Siontis, G.C., Tzoulaki, I., Ioannidis, J.P.: Predicting death: an empirical evaluation of predictive tools for mortality. *Arch. Intern. Med.* **171**(19), 1721–1726 (2011)
20. Song, H., Rajan, D., Thiagarajan, J.J., Spanias, A.: Attend and diagnose: clinical time series analysis using attention models. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
21. Trask, A., Gilmore, D., Russell, M.: Modeling order in neural word embeddings at scale. arXiv preprint [arXiv:1506.02338](https://arxiv.org/abs/1506.02338) (2015)
22. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
23. Wang, Y., Kung, L., Byrd, T.A.: Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018)