Information Sciences 637 (2023) 118984

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Dual-core mutual learning between scoring systems and clinical features for ICU mortality prediction

Zhenkun Shi [a,b,c,*], Sen Wang [c], Lin Yue [d], Yijia Zhang [e,*], Binod Kumar Adhikari [b], Shuai Xue [f], Wanli Zuo [b,*], Xue Li [c]

[a] *Tianjin Institutes of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China*
[b] *College of Computer Science and Technology, Jilin University, Jilin 130000, China*
[c] *School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane 4072, Australia*
[d] *School of Information and Physical Sciences, The University of Newcastle, Newcastle 2308, Australia*
[e] *College of Electronic Countermeasures, National University of Defense Technology, Hefei 230027, China*
[f] *The First Hospital of Jilin University Changchun, Jilin 130021, China*

ARTICLE INFO

Keywords:
Health informatics
Mutual learning
Scoring system
Mortality prediction

ABSTRACT

Perpetually improving mortality prediction in intensive care units (ICUs) via the implementation of eHealth evaluation approaches has become a major research hotspot in the field of medical data mining for the purpose of saving lives. Recently, researchers have attempted to achieve improved prediction accuracy by using only deep learning-based techniques. However, some problems remain. (1) Most of the existing methods utilize independent clinical features to predict mortality by eliminating the correlations between the latent features, but this technique may fail to comprehensively capture and evaluate the statuses of patients. (2) Several clinical features are needed to ensure strong prediction accuracy, but most methods only use static features that are not extendable. (3) An effective practical framework that unifies traditional ICU scoring systems and state-of-the-art deep learning methods to predict mortality is also lacking. (4) Moreover, the interpretability of existing deep learning-based methods needs to be further improved. Therefore, we propose a novel **d**ual-core **m**utual **l**earning **f**ramework (DMLF) between ICU scoring systems and clinical features for mortality prediction. In particular, we mutually utilize **s**equential **o**rgan **f**ailure **a**ssessment (SOFA) scores and clinical measurement features to learn a unified model for enhancing the accuracy and interpretability of our DMLF. Experiments conducted on five real-world disease datasets show that the DMLF achieves significantly better prediction accuracy and area under the receiver operating characteristic curve (AUROC) values than six baselines and four state-of-the-art methods. Moreover, clinicians utilize a familiarized SOFA system to conduct mortality prediction and achieve increased interpretability, which benefits the adoption of the proposed framework in real clinical scenarios.

* Corresponding author.
E-mail addresses: shizk14@mails.jlu.edu.cn (Z. Shi), zhangyj_gfkj@163.com (Y. Zhang), zuowl@jlu.edu.cn (W. Zuo).

https://doi.org/10.1016/j.ins.2023.118984
Received 17 January 2022; Received in revised form 16 February 2023; Accepted 17 April 2023

Available online 20 April 2023

Table 1
Mortality prediction performance measures achieved on the MIMIC III dataset using ICU scoring systems.

CCS	SOFA			SAPS II		
	AUC	PRC	ACC	AUC	PRC	ACC
2	0.6944	0.4791	0.7368	0.76	0.5706	0.7438
49	0.7051	0.316	0.8681	0.7975	0.4245	0.8508
98	0.6622	0.2507	0.888	0.7991	0.3837	0.8727
101	0.6798	0.2379	0.8917	0.7955	0.3508	0.8647
157	0.6993	0.4413	0.7697	0.7592	0.525	0.7606

CCS: Clinical Classifications Software (CCS) for ICD-9-CM [15];
2: Septicemia (except in labor);
49: Diabetes mellitus without complications; 98: Essential hypertension;
101: Coronary atherosclerosis and other heart disease;
157: Acute and unspecified renal failure.

1. Introduction

Mortality prediction is one of the major research areas in the field of medical data mining; its goal is to provide accurate death rate assessments. Mortality prediction is extremely vital in intensive care units (ICUs) because ICU clinicians need to react quickly when making decisions and taking actions based on the estimated mortalities to save lives. Moreover, mortality prediction is promising for effectively improving health management and reducing the rate of needless deaths [1]. Owing to the demand for mortality prediction, a rising number of studies [2–6] have been successfully completed from numerous viewpoints, such as how to predict mortality based on the diagnosis of diseases [2], how to model clinical time series data [3,4,7], and how to integrate domain knowledge into mortality prediction [5,6].

With the massive increase in the number of electronic health records (EHRs), clinical data have been extensively collected and used in precision medicine. Hopefully, a large amount of real-world clinical data are publicly available for researchers, thereby boosting studies that use EHRs for clinical data mining. Conventional methods, such as the chronic health evaluation (APACHE) [8], sepsis-related organ failure assessment (SOFA) [9], and simplified acute physiology score (SAPS) [10], have implemented clinical scoring systems in clinical practice. These scoring systems were established based on physiological data and clinical rules. The advantages of scoring systems in real clinical scenarios are their ease of implementation and high interpretability. However, the limitation of such scoring systems is their low accuracy because these scoring systems use limited indicators and fixed clinical rules.

Recently, advancements in deep learning have made it possible to more accurately and effectively predict mortality. Researchers have made many efforts to use deep learning techniques for mortality prediction [3,2,6,11–13]. Shi et al. attempted to predict mortality by using an attention mechanism and a multitask technique [3,2]. Liu et al. constructed a deep dual network integrated with medical domain knowledge to predict mortality. Suresh et al. proposed a benchmark for mortality prediction and other learning tasks involving ICU data mining [11,14]. Zhang et al. tried to solve mortality prediction with limited EHRs by introducing a meta-learning method. However, all of these methods were explicitly designed to seek more accurate predictions, but they fail to comply with the clinical procedures and information contained in the traditional information comprehension methods. Hence, the existing scoring systems are not highly useful in medical knowledge cases. Moreover, deep learning-based methods usually need large amounts of data to ensure their prediction quality. Furthermore, deep learning techniques are often described as black boxes; hence, their interpretability is limited, which narrows their applicability in real-world clinical scenarios. In addition, the features used in these methods are fixed, so their stability and robustness are not sufficiently reliable.

To tackle these limitations, we propose a novel **d**ual-core **m**utual **l**earning **f**ramework (DMLF) between ICU scoring systems and clinical features for mortality prediction. The DMLF can utilize the advantages of existing clinical scoring systems and deep learning techniques to conduct mortality prediction while achieving high accuracy and reasonable interpretability.

1.1. Motivations and objectives

To bridge the gap between the existing ICU scoring systems and state-of-the-art deep learning methods, we propose two objectives to achieve better mortality predictions with higher interpretability.

Objective 1: To design an ICU mortality prediction learning framework with high performance and preserve its interpretability by utilizing existing scoring systems. As shown in Table 1, ICU scoring systems achieve low performance when predicting mortality (*e.g.*, the SOFA has a 34.5% area under the precision-recall curve (AUPRC) and a 68.18% area under the receiver operating characteristic curve (AUROC)). We consider the recent advances in mortality prediction by using deep learning frameworks [3,11]. We note that the integration of a scoring system into a learning framework can not only yield better performance but also preserve the structure of the scoring system. The benefit of this framework is that it achieves more reliable interpretability, enabling it to adapt to real clinical circumstances. Therefore, we implement the SOFA as the scoring system for building the learning framework in this paper.

Objective 2: To achieve state-of-the-art prediction performance through mutual learning mechanisms. Inspired by the success of related works on mortality prediction [3,16–18] and the power of mutual learning [19], Shi et al. [3] achieved nearly 97.89% accuracy and a 93.69% AUROC on diseases concerning essential hypertension by using EHRs from the MIMIC III dataset

[20]. We believe that it is possible to achieve an enhanced learning performance by utilizing mutual information derived from a traditional scoring system and a state-of-the-art learning method. As the employed scoring system is well designed and has been tested in practical situations for a long time, it can also be effectively implemented from a theoretical perspective. Moreover, the most significant advantage of a scoring system is its interpretability. Learning methods are designed to learn latent representations from unhandled raw EHRs, and experiments on real-world datasets demonstrate the correctness and high performance of these methods. Therefore, theoretically, we utilize the mutual learning among these kinds of methods to achieve both improved interpretability and enhanced performance. Therefore, the second goal in this work is to build a unified dual-core framework to mutually learn latent information from scoring systems and raw clinical features.

1.2. Our contributions

To eliminate the deficiencies of existing techniques, first, we present a dual-core mutual learning framework to learn from ICU scoring systems and raw clinical EHRs; then patients' severity levels are evaluated in this work. We also formulate the obtained SOFA scores as a directed graph $G = (V, E)$ with a node set V and an edge set E. V stands for the SOFA score, which indicates the patient's current severity status, and E represents the status change trajectory. Second, we learn latent representations from the raw EHRs by using residual gated recurrent units (RGRUs). Next, we mutually use an attention mechanism and a cross-shared unit (CSU) to learn from the SOFA graph and the raw clinical features. Finally, we use a unified state layer to evaluate mortality and predict the results. Our contributions are summarized as follows. 1) The SOFA scores are formulated as a gated graph neural network (GGNN), which is integrated into a unified learning framework. 2) RGRUs are introduced to learn latent representations from the unhandled raw clinical EHRs. 3) A CSU is explicitly designed to mutually leverage the information derived from the GGNN and RGRUs and establish a unified framework (the DMLF). 4) The effectiveness of our proposed framework is verified on the top 5 most frequently diagnosed diseases from the real-world MIMIC-III dataset. The experiments demonstrate that our method provides the best performance in comparison with a variety of baseline methods and five state-of-the-art methods.

2. Related work

In general, two categories of approaches are available: case-oriented methods with the medical domain and statistical knowledge-based methods [21]; these approaches are called traditional methods (or data-driven methods with data mining techniques) and deep learning algorithms (or deep methods), respectively. We briefly review some of the relevant works in each category and refer readers to [22,23] for comprehensive surveys.

Traditional methods. The earliest approach for evaluating patient severity is the life table [21], which tries to use fixed physiological features to assess mortality. Later, a variety of scoring systems were developed to evaluate severity, such as APACHE scores [8], Glasgow coma scale (GCS) [24] scores, SAPSs [10] and SOFA [9] scores. In recent years, researchers have attempted to adopt machine learning techniques, such as support vector machines (SVMs), logistic regression (LR), extreme gradient boosting (XGBoost), and random forests (RFs) [25], to solve this problem with the use of EHRs. The traditional mortality prediction methods utilize EHRs [26,6] and usually aggregate clinical features first and then make predictions, but they ignore the temporal information gap between features and the latent information among feature sequences. Although most of these traditional methods have the capability to provide clinical interpretations, their applications are constrained due to inaccuracy.

Deep methods. In recent years, deep learning approaches have achieved exceptional success in many areas by constructing deep hierarchical features and capturing longitudinal dependencies within data [25]. The supremacy of deep methods is due to their better performance and reduced use of feature engineering [27].

BK-DDN [6] is a knowledge-aware deep dual network for mortality prediction that fuses the representations of medical knowledge and raw text for prediction and achieves good prediction performance. Suresh et al. [11] proposed a benchmark task for mortality prediction and submitted a baseline by using multitask learning. However, the article [6,11] failed to consider the local temporal dependencies in EHRs and information about traditional scoring systems. To capture local and temporal dependencies, Che et al. [28] proposed a multilayer convolutional neural network (CNN) model to learn medical feature embeddings to address the problems of high dimensionality and temporality for mortality prediction [28]. However, this model fails to establish the latent relations among different features.

Recently, the recurrent neural network (RNN) and its variants have demonstrated the importance of handling time series data. Many efforts have been made to model EHRs by utilizing RNNs, and these approaches have achieved the best performance in terms of the accuracy and area under the curve (AUC) metrics [27,29–31]. The most obvious drawback of employing RNNs is their interpretation difficulty. This may not create a problem in some areas (e.g., image classification) because the end user can distinguish whether the obtained result is true or not. However, RNNs are not applicable in the medical domain because medical results cannot be intuitively judged, and it is also difficult for clinicians to trust the results of a model with weak interpretability. To enhance the ingenuity of interpretation, graph-based models have been developed to solve learning problems, and they have achieved great success in natural language processing (NLP) [32], computer vision (CV) [33], and social computing [34]. However, some issues remain regarding the transplantation of graph models into the healthcare area. The first concern is the node edge selection problem. In this case, we cannot directly select all the records as nodes because the number of records is massive. Therefore, it is impossible to directly design such a large graph. The second issue is how to model temporal information in a graph.

In this work, to bridge the gap between traditional scoring systems and the most advanced deep learning techniques, we build a DMLF to predict mortality for ICU patients. The first core is a SOFA sequential graph core (SSGC), which is a sequential SOFA

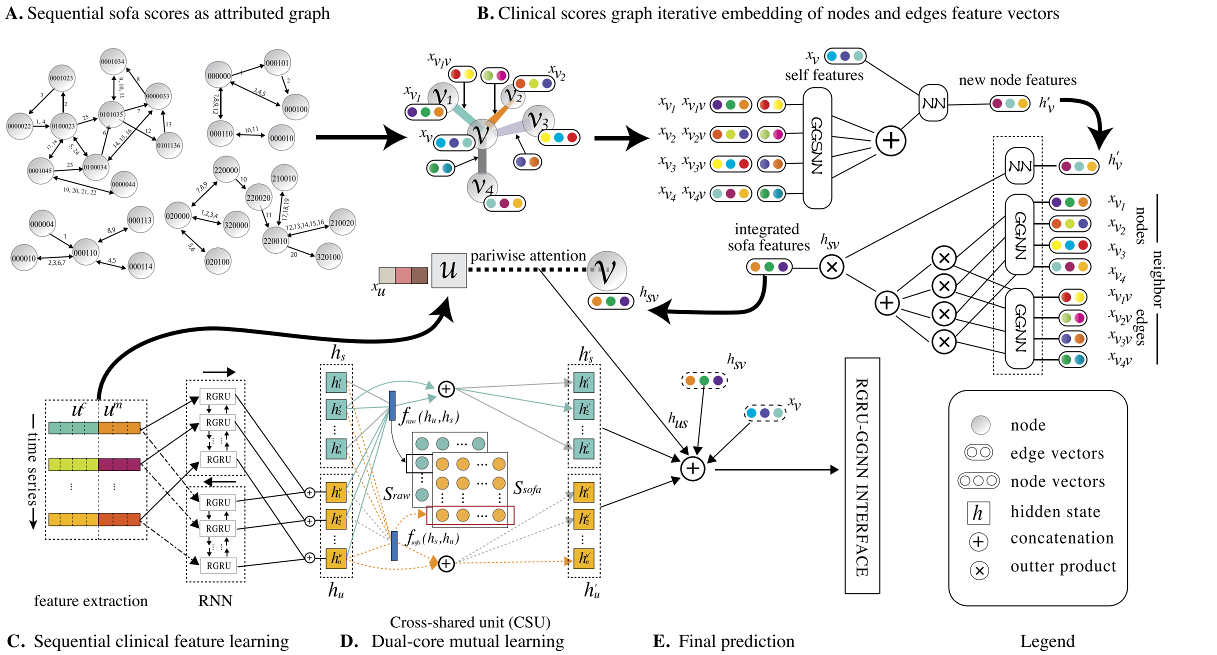


Fig. 1. The DMLF. **(A)**: We start from a patient's SOFA score graph and **(C)** a raw time series clinical feature matrix representation of their clinical measurements and medical treatments. **(B)**: By incorporating neighboring node information, we iteratively update the node feature vectors. By applying multiple iterations of node embedding, based on the representation of the node of interest and those of the node's neighbors, a neighboring SOFA feature vector h_s is calculated. To account for the effects between raw clinical features and SOFA scores, a CSU is designed for the mutual learning **(D)** of latent representations from the hidden states of the SOFA scores and raw clinical features. In addition, a pairwise attention mechanism is adopted to produce a context vector for each SOFA node and raw clinical feature pair $< v, r >$ as a learned, weighted combination of all sequential nodes and raw feature vectors. Finally, **(E)** through concatenation, the SOFA features, context vectors, and latent representation vectors from the CSU are used to jointly predict mortality.

score-constructed graph neural network. The second core is a raw data learning core (RDLC), which is a raw EHR neural network created by utilizing RGRUs. We mutually acquire the information from the SSGC and RDLC and use it for mortality prediction. The DMLF achieves the best prediction performance and preserves the interpretability derived from the scoring system. The problem statement and our methodology are discussed in the next section.

3. Methodology

The use of supervised deep learning techniques for clinical trials has enabled a new level of prediction performance beyond that of the traditional existing clinical methods (i.e., clinical scoring systems). However, high-performance approaches bring trustworthiness concerns due to the "black box" problems in deep learning methods, so clinicians have historically distrusted neural network models [35]. To achieve our first objective, a novel mortality prediction framework is designed by using an existing ICU scoring system. In this work, we use the SOFA scoring system for Modeling purposes. Scoring systems have been adopted in real clinical practice for many years and have proven their efficacy. Clinicians are familiar with these systems and know their principles and internal mechanisms well. Scoring systems are utilized and integrated into neural networks to conduct mortality prediction, which will improve both prediction performance and model interpretability. The architecture of the SOFA scoring system model is presented in Fig. 1 **(A)** and Fig. 1 **(B)**.

To satisfy our second objective, we implement a state-of-the-art prediction approach to achieve improved performance. It is inadequate to solely utilize clinical scoring systems to obtain the best performance because scoring systems use a relatively small amount of clinical features to indicate a patient's severity. For example, SOFA uses ten clinical indicators: partial pressure of oxygen, fraction of inspiration oxygen, platelets, bilirubin, mean arterial pressure, dobutamine, dopamine, Glasgow coma score, creatinine, and urine output. Upon comparing hundreds of monitoring and laboratory testing indicators that inadequately use scoring systems to comprehensively evaluate a patient's status, undoubtedly, neural networks cannot achieve the best performance through the sole use of these ten indicators. In this work, we incorporated more clinical indicators derived from raw clinical EHRs into the neural network to boost its prediction performance. The method of introducing raw clinical features into the main framework to boost performance is shown in Fig. 1 **(C)** and Fig. 1 **(D)**.

3.1. Problem statement

For T given hours of an ICU stay, it is assumed that a series of clinical observations $X = \{x_1, x_2, \cdots, x_i\}_{i\geq 1}^{T}$ have been made during the stay, where $x_i = [x_{is}; x_{ir}]$ is a concatenated vector that represents the i-th observation vector. x_{is} represents the SOFA score vector (the scores are based on six different aspects: cardiovascular, hepatic, coagulation, renal and neurological systems.)[1] in the i-th observation, and x_{ir} represents the raw clinical feature vector (e.g., temperature, sodium, glucose, heart rate, etc.) in the i-th observation. For each ICU stay, we also provide a binary label $y \in \{0, 1\}$ to indicate whether the patient died within a specific time period. Our objective is to generate a sequence-level mortality prediction for each ICU stay.

3.2. SOFA score graph modeling

Let us define $G = (V, E)$ as a directed graph, where V is the set of nodes and E is the set of edges. Node $v \in V$ takes unique values from the SOFA scores $1, 2, \cdots, |V|$ (SOFA records six systems with scores ranging from 0-4, and the number of nodes is $5^6 = 15625$), and the edges are paired as $e = (v, v') \in V \times V$. The node vector for node v is denoted by $h_v \in \mathbb{R}^D$. We add node labels (the patient's physiological features that are used to calculate their SOFA scores) $l_v \in 1, 2, \cdots, L_v$ for each node v and edge labels (diagnosed patient diseases) $l_e \in 1, 2, \cdots, L_e$ for each edge. We overload the notations as $h_s = \{h_v | v \in S\}$ when S is a set of nodes and as $l_s = \{l_e | e \in S\}$ when S is a set of edges. We use the $IN(V)$ function and $OUT(V)$ function to obtain the predecessor node set and successor node set, where $IN(V) = \{v' | (v', v) \in E\}$ and $OUT(V) = \{v' | (v, v') \in E\}$ return node sets v' with $v' \to v$ and $v \to v'$, respectively. The neighbor set of node V is denoted as $NB(v) = IN(V) \cup OUT(V)$, and the neighbor edge set is denoted as $NBE(v) = \{(v', v'') \in E | v'' = v' \vee v\}$.

As shown in Fig. 1 **(B)**, we map the SOFA graph to produce an output via two steps. First, we compute the node representation for each node via interactive embedding, namely, the propagation step. Next, for each $v \in V$, we use the output model $O_v = g(h_v, l_v)$ to map the node representations and correspondingly label them to outputs o_v. The initial node representation h_v^0 is set to x_v, where x_v is the input node. Then, each node representation is updated via the update function f^*:

$$\boldsymbol{h}_v = f^*(l_v, l_e, l_{NBE_v}, (h^1_{NBE_v}, \cdots, h^i_{NBE_v})), i \geq 1 \tag{1}$$

where f^* is a neural network and i is a predefined parameter used to control the spread breadth. We formulate the final representation of node v as:

$$\boldsymbol{h}_{sv} = \sum_{j \in NBE_v} \boldsymbol{a}_j \boldsymbol{h}_{v_j}, \sum_{j \in NBE_v} a_j = 1, a_j \geq 0 \text{ for } j \in NEB_v \tag{2}$$

where $\boldsymbol{h}_{sv} \in \mathbb{R}^m$ denotes the final representation of SOFA node v, NBE_v represents the neighbors of node v, h_{v_j} indicates the basic embedding of node v_j, and $a_j \in \mathbb{R}^+$ is the attention weight on embedding h_{v_j} when $\boldsymbol{h}_{sv}$ is calculated. The attention mechanism is discussed in Section 3.6.

3.3. Propagation model

Similar to the work of Li et al. [36] and Beck et al. [32], we unroll the recurrence and use backpropagation to compute the gradients. Unlike in their works, where the recurrence was unrolled to a fixed number of steps, we define the number of unrolling steps as a dynamic value in this work. The RNN used in our work is a GRU, and its full component, the GGNN, is defined as follows:

$$\boldsymbol{r}^t_v = \sigma(c^r_v \sum_{v' \in NBE_v} \boldsymbol{W}^r_{l_e} h^{t-1}_{v'} + \boldsymbol{b}^r_{l_e}) \tag{3}$$

$$\boldsymbol{z}^t_v = \sigma(c^z_v \sum_{v' \in NBE_v} \boldsymbol{W}^z_{l_e} h^{t-1}_{v'} + \boldsymbol{b}^z_{l_e}) \tag{4}$$

$$\tilde{\boldsymbol{h}}^t_v = \rho(c_v \sum_{v' \in NBE_v} \boldsymbol{W}_{l_e}(r^t_{v'} \odot h^{t-1}_{v'}) + \boldsymbol{b}_{l_e}) \tag{5}$$

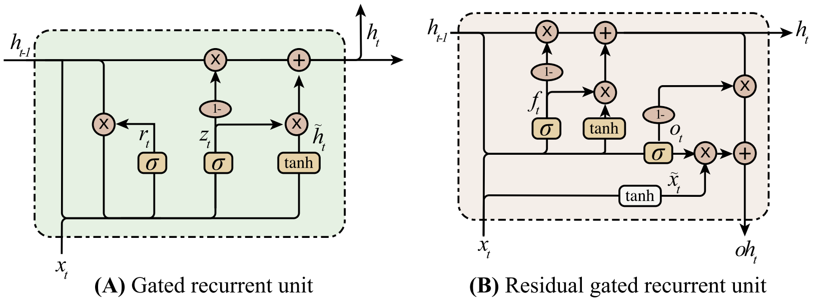
$$\boldsymbol{h}^t_v = (1 - z^t_v) \odot h^{i-1}_v + z^t_v \odot \tilde{\boldsymbol{h}}^t_v \tag{6}$$

where σ is the sigmoid function and ρ is a nonlinear function. We initialize $\boldsymbol{h}^0_v = x_v$, where x_v indicates the embedded representation of node v. To eliminate the scale factors that might exist between variables, we add normalization constants to the control gates and hidden state, where $c_v = c^z_v = c^r_v = |N_v|^{-1}$.

3.4. Sequential clinical feature learning by utilizing raw EHRs

Raw clinical EHRs carry exhaustive information, which helps to comprehensively evaluate a patient's severity. In this work, we utilize raw clinical features to boost the learning performance of our approach. Raw clinical features are composed of two parts: clinical charting features u_c and medical treatments features u_m. Instead of simply concatenating the raw clinical features, we first use the window alignment operation proposed in Shi et al. (2019) [2] to simultaneously align the clinical measurements and medical treatments in step t. Then, we use a stacked RNN to learn representations h_c and h_m for the clinical measurements and medical treatments, respectively. Next, we concatenate $h_u = h_c \odot h_m$, the raw clinical representations, as the input of the neural network. $\odot$ represents the concatenation operation. u_c and u_m denote the raw clinical features obtained from clinical measurement and the

[1] https://en.wikipedia.org/wiki/SOFA_score.



(A) Gated recurrent unit

(B) Residual gated recurrent unit

Fig. 2. GRU and RGRU.

medical treatment features in Fig. 1 **C**, respectively. All time series data u_c and u_m are sliced according to their SOFA scores during the same time span. The length of each ICU stay is cut or padded with zeros to a fixed number of hours, and an observation is made once per hour. The main architecture of the raw clinical feature learning process is a stacked RGRU. We introduce the RGRU in detail in the following section.

3.5. Stacked dual RGRU

A stacked RNN layer is designed to capture the dependencies between the raw clinical features and patients' severity levels, and the RNN layer is realized by an RGRU component. As shown in Fig. 2, similar to the traditional GRU, the RGRU also has two gates to control the flow of the input and hidden states. However, their implementation mechanisms are different. Given an input x_t at time t and the previous state h_{t-1}, the new state h_t is calculated by the following equations:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{7}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{8}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{f}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{f}_t \odot tanh(\boldsymbol{W}_{xi}x_t) + \boldsymbol{b}_h \tag{9}$$

$$\boldsymbol{oh}_t = (1 - \boldsymbol{o}_t) \odot \boldsymbol{h}_t + \boldsymbol{o}_t \odot \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_{oh} \tag{10}$$

where f_t is a forget gate used to control the information flow from the previous time step $t-1$ to the current time step t and o_t is a residual gate employed to control the information flow from the previous layer to the next layer. o_t is a residual gate that controls the information flow from the previous layer to the next layer. σ is a sigmoid function, and $\odot$ denotes elementwise multiplication. $\boldsymbol{W}_*$ is a weighted parameter metric, where $* \in \{xf, hf, xo, ho, i, xoh\}$. $\boldsymbol{b}_*$ is the bias vector, where $* \in \{f, o, h, oh\}$. $\tilde{\boldsymbol{x}}_t$ is calculated by the following function:

$$\tilde{\boldsymbol{x}}_t = \begin{cases} x_t & \text{if } size(x_t) = size(h_t), \\ tanh(\boldsymbol{W}_{\boldsymbol{xoh}}\boldsymbol{x}_t) & \text{otherwise.} \end{cases} \tag{11}$$

EHRs have strong time dependencies. For example, clinical measurements (e.g., body temperatures) taken at t_i affect medical treatments (e.g., aspirin) performed at t_j, where $i < j$. Similarly, medical treatments (e.g., insulin) performed at t_j can also reflect a patient's physiological status (e.g., blood glucose) at t_i, where $i < j$. This is very similar to the word dependencies in a sentence in the field of NLP [37]. To capture these time dependencies, a bidirectional RGRU (BRGRU) is adopted in this work. The BRGRU uses two directional representations for the time series inputs, making it similar to a bidirectional GRU [38]. As shown in Fig. 1 **C**, we concatenate the hidden states generated by the RGRU in both directions by using the following equations:

$$\boldsymbol{h_t} = \overleftarrow{\boldsymbol{h_t}} \oplus \overrightarrow{\boldsymbol{h_t}} \tag{12}$$

where $\overrightarrow{\boldsymbol{h_t}}$ is the clockwise direction, and $\overleftarrow{\boldsymbol{h_t}}$ is the counterclockwise direction.

3.6. Mutual attention learning from SOFA scores and raw clinical features

Considering the interrelationships among the SOFA scores and the outer relationships between the clinical features and SOFA scores, as shown in Fig. 1 **(D)**, we add two attention layers. The first is a SOFA score self-attention layer that calculates the weights of SOFA scores. The second is a pairwise mutual attention layer that reveals the correlation between the SOFA scores and raw clinical features. As shown in Fig. 1 **(B)**, in a SOFA graph, each node v_i is calculated with an integrated SOFA vector $h_{sv} \in \mathbb{R}^m$, where m represents the embedding dimensionality. Then, $h_{s1}, h_{s2}, \cdots, h_{s|V|}$ are the SOFA embeddings of nodes $v_1, v_2, \cdots, v_{|V|}$. The SOFA node embedding calculation process is described in Section 3.2. The attention weights are calculated by using a softmax function, as shown in Eq. (2):

$$\boldsymbol{a}_j = \frac{exp(f(\mathbf{h}_v, \mathbf{h}_{v_j}))}{\sum_{k \in NBE_v} exp(f(\mathbf{h}_v, \mathbf{h}_{v_k}))} \tag{13}$$

where $f(h_v, h_{v_j})$ is a scalar value that represents the compatibility between the basic embeddings $\mathbf{h}_v$ and $\mathbf{h}_{vj}$. Similar to the work of Edward et al. [39], we use a feedforward network and a single hidden layer to compute $f(h_v, h_{v_j})$:

$$f(h_v, h_{v_j}) = q_a^\top tanh(\boldsymbol{W}_a \begin{bmatrix} h_v \\ h_{v_j} \end{bmatrix} + b_a) \tag{14}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{n \times 2m}$ is a weight matrix for the concatenated vector $\boldsymbol{h}_{v \to j}$ and $\boldsymbol{b}_a$ is a bias vector for generating the scalar values. n represents the dimensions of the hidden state $f(\cdot, \cdot)$.

Different from SOFA score self-attention, we use scaled dot-product attention [40] to calculate the pairwise mutual attention mechanism. To implement this mechanism, we first generate a representation of $\boldsymbol{u}^c$ and $\boldsymbol{u}^m$. We denote this representation as $\boldsymbol{h}_u = f(\boldsymbol{u}^c, \boldsymbol{u}^m)$, where $f(\cdot, \cdot)$ is a simple embedding function. The hidden representation of a SOFA score $\boldsymbol{h}_{sv}$ is calculated in Section 3.2. Then, we use $\boldsymbol{h}_u$ and $\boldsymbol{h}_{sv}$ to conduct an attention concatenation operation. The attention score calculation process can be expressed as follows:

$$Attention(\boldsymbol{Q}, (\boldsymbol{K}, \boldsymbol{V})) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}\top}{\sqrt{d}})\boldsymbol{V} \tag{15}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are the matrices formed by the query, key, and value vectors, respectively, and d is the dimensionality of the key vectors. We use $\boldsymbol{h}_u$ and $\boldsymbol{h}_{sv}$ as the query vector and tuple $(key, value)$ vector for alternately calculating the attention scores λ_u, λ_{sv}. As shown in Fig. 1 **(D)**, we concatenate $\boldsymbol{h}_{us} = \lambda_u \boldsymbol{h}_u \oplus \lambda_{sv} \boldsymbol{h}_{sv}$ as one of the inputs for participating in the final predictions.

3.7. Dual-core mutual learning via a CSU

After the representations of the GGNN and RGRU are obtained, the pieces of information derived from the traditional SOFA system and the raw EHRs are separated from each other. However, the SOFA scores and raw clinical features have strong relations regarding the assertion of patient severity. For instance, the SOFA scores from the respiratory system are mainly based on PaO_2/FiO_2, and PaO_2/FiO_2 is closely related to other features, such as the respiratory rate, heart rate, O_2 flow and tidal volume. Compared to the SOFA scores, the raw records can provide more information for asserting severity, but they provide smaller granularity. The SOFA system was explicitly designed by clinicians and experts and has been tested empirically, and its efficacy in asserting severity is moving in a proper direction. SOFA scores are used to evaluate patient mortality, and they are more intuitive than the raw figures. In addition, the SOFA system is also used for interpretability.

The CSU is designed to consider the interactions between raw EHRs and the SOFA system. Let U, u and S, s represent the indices of the raw clinical feature embeddings and the SOFA feature embeddings, respectively. We overload the notation to assume the following:

$$\begin{cases} \bar{m} = s, M = U & \text{if } m = a \\ \bar{m} = u, M = S & \text{if } m = u \end{cases} \tag{16}$$

where $M \in \{U, S\}$ and $m \in \{u, s\}$. We first compute the composition vector $\boldsymbol{a}_{ij}^{\boldsymbol{M}} \in \mathbb{R}^{\boldsymbol{N}}$ through the following tensor operator: $a = up = s$

$$\boldsymbol{a}_{ij} = f_m(\boldsymbol{h}_i^m, \boldsymbol{h}_j^{\bar{m}}) = tanh((\boldsymbol{h}_i^m)\top \boldsymbol{G}^m \boldsymbol{h}_j^{\bar{m}}) \tag{17}$$

where $\boldsymbol{h}_i^m \in \boldsymbol{h}_M$ is the hidden representation at the i-th time step, and $\boldsymbol{G}^m \in \mathbb{R}^{N \times 2d \times 2d}$ are 3-dimensional tensors. $\boldsymbol{N}$ is a unified hyperparameter. These tensor operations can be seen as multiple bilinear terms that have the capability of modeling more complicated compositions between two vectors [41,37]. After obtaining the composition vectors, a cross-shared score $\boldsymbol{SH}_{ij}^M$ is calculated by the following equation:

$$\boldsymbol{SH}_{ij}^M = \boldsymbol{v}_m^\top \boldsymbol{a}_{ij}^M \tag{18}$$

where $\boldsymbol{v}_m \in \mathbb{R}^N$ is a weight vector used to weight each value of the composition vector. $\boldsymbol{SH}_{ij}^M$ can be regarded as a scalar. As shown in Fig. 1 **(D)**, we use two matrices $\boldsymbol{S}_{raw}$ and $\boldsymbol{S}_{sofa}$ to record these scalars. A higher score $\boldsymbol{SH}_{ij}^U$ indicates a higher correlation between the i-th clinical feature and the j-th system of the human body (the SOFA scoring system includes six human body systems). Similarly, a higher score $\boldsymbol{SH}_{ij}^S$ indicates a higher correlation between the i-th human body system and the j-th raw clinical features.

3.8. Unified output

A unified graph-level output mechanism is implemented for mortality prediction. First, we generate a node level representation by using the hidden state and attention scores with the help of the following calculations:

$$\boldsymbol{h}_v = f^g(\boldsymbol{x}_v, \boldsymbol{h}_{sv}, \boldsymbol{h}'_s, \boldsymbol{h}'_u, \boldsymbol{h}_{us}) \tag{19}$$

Table 2
Detailed description of the dataset utilized for the mortality prediction task.

CCS ID	CCS Name	ICD9 Code	Samples	Nodes	Edges	Death Rate
2	Septicemia (except in labor)	0031, 0202, 0223, 0362, 0380, 449, 0381, 0383, 03811, 03812, 03819, 0382, 03810, 03840, 03841, 03843, 03844, 03849, 0388, 0389, 0545, 77181, 7907, 99591, 99592, 03842,	1,3426	6,340	49,904	0.1496
49	Diabetes mellitus without complications	24900, 7915, 25001, 7902, 79021, 79022, 79029, 25000, 7916, V4585, V5391, V6546	1,2808	4,024	27,342	0.0889
98	Essential hypertension	4011, 4019	2,5851	4,539	31,835	0.0810
101	Coronary atherosclerosis and other heart diseases	4110, 4111, 4118, 41181, 41189, 4148, 4130, 4131, 4139, 4140, 4149, 41401, 41406, 4142, 4143, 4144, 412, 41400, V4581, V4582	1,9315	4,230	30,966	0.0740
157	Acute and unspecified renal failures	5845, 5846, 5847, 5848, 5849, 586	1,7522	6,575	56,213	0.1387

Table 3
Clinical feature list extracted from the raw EHRs.

CATEGORY	NAME
CLINICAL MEASUREMENTS	1. Age, 2. Admission Location, 3. Admission Type, 4. Current Service, 5. Ethnicity, 6. Gender, 7. Height, 8. Religion, 9. Weight 10. GCS Total, 11. GCS Verbal, 12. GCS Motor, 13. GCS Eyes, 14. Arterial BP [Diastolic], 15. Arterial BP [Systolic], 16. Arterial BP [Mean], 17. Manual BP [Diastolic], 18. Manual BP [Systolic], 19. NBP [Diastolic], 20. NBP [Systolic] 21. Base Excess, 22. Chloride, 23. Calcium, 24. FiO2, 25. FiO2 Set, 26. Heart Rate, 27. Heart Rhythm, 28. Mean Airway Pressure 29. O2 Flow, 30. PCO2, 31. PO2, 32. Respiratory Rate, 33. Respiratory Rate Set, 34. SpO2, 35. Temperature, 36. Total CO2, 37. Urine Output 38. Bicarbonate, 39. Carboxyhemoglobin, 40. Glucose, 41. Hematocrit, 42. Hemoglobin, 43. Intubated, 44. Lactate, 45. Methemoglobin, 46. O2 Flow, 47. PH, 48. SO2, 49. Sodium, 50. Tidal Volume, 51. Ventilation Rate, 52. Ventilator 53. Albumin, 54. Bilirubin, 55. Blood Urea Nitrogen, 56. Creatinine, 57. Magnesium Sulfate, 58. Platelet, 59. Potassium, 60. WBC
MEDICAL TREATMENTS	1. Dextrose (5%) in Water, 2. NaCl, 3. Propofol, 4. Insulin, 5. Heparin, 6. Fentanyl, 7. Neosynephrine, 8. Phenylephrine, 9. Midazolam, 10. Amiodarone, 11. Dopamine, 12. Potassium Chloride, 13. Vasopressin, 14. Triphospho Pyridine Nucleotide, 15. Nitroglycerin, 16. Piperacillin, 17. Milrinone, 18. Nitroprusside, 19. Morphine Sulfate, 20. Epinephrine, 21. Dobutamine, 22. Diltiazem, 23. Cisatracurium, 24. Calcium Gluconate, 25. Dilaudid, 26. Sodium Bicarbonate, 27. Nicardipine, 28. Esmolol, 29. Labetalol, 30. Peptamen, 31. Lidocaine, 32. Thiamine, 33. Metoprolol, 34. Piperacillin, 35. Ampicillin, 36. Albumin 25%

where $\boldsymbol{x}_v$ is the initial node input, $\boldsymbol{h}_s v$ is the node representation derived from its neighbors, $\boldsymbol{h}'_s, \boldsymbol{h}'_u$ is the node representation from the CSU, and $\boldsymbol{h}_{us}$ is the node representation obtained from the raw clinical features and SOFA scores. Then, we define a graph-level representation vector as:

$$\boldsymbol{h}_g = Softmax(\sum_{v \in V} \sigma(NN_1(\boldsymbol{h}_v^\top, \boldsymbol{x}_v) \odot \tanh(NN_2(\boldsymbol{h}_v^\top, \boldsymbol{x}_v))) \tag{20}$$

where $\sigma(NN_1(\boldsymbol{h}_v^\top, \boldsymbol{y}_v)$ acts as a soft attention mechanism that decides which nodes are relevant to the current graph-level prediction. NN_1 and NN_2 are neural networks that take the concatenation of $\boldsymbol{h}_v^\top$. $\boldsymbol{x}_v$ is the graph input that outputs real-valued vectors.

4. Experiments

4.1. Data description

We conducted experiments on a real-world, publicly available, and deidentified dataset called MIMIC III [20]. It contains structured (e.g., real-time sensor data, laboratory tests, and treatments) as well as unstructured data (e.g., free-text clinical notes) for more than 60,000 ICU admissions between 2001 and 2012 (mainly from two US hospitals).

The dataset was processed to select a cohort of patients who provided meaningful evaluations of mortality prediction methods. We filtered the patients whose ICU stay lengths were less than one hour or whose ages were below 16. Based on the patients' diagnosis results listed in MIMIC III, we used clinical classification software to group the patients. We selected the top 5 most frequently diagnosed diseases as the data for our experiments. The dataset details are shown in Table 2. In our work, the data were randomly split into a training set, validation set, and testing set using a $70\% : 20\% : 10\%$ ratio.

For each step, we used a six-dimensional vector ($x \in \mathbb{R}^6$) to record a SOFA score; each dimension represented one body system for the SOFA score system. In this work, we used two groups of raw clinical data: clinical measurement data and medical treatment data. We selected the top 70 most frequently used clinical measurement features and the top 35 most frequently used medical treatment features. We list all these features in Table 3.

4.2. Experimental settings

In our experiment, we collected an observation every hour after a patient was admitted to the ICU to set a time gap of one hour and made a mortality prediction for 48 hours after each patient's admission. The dimensions of $\boldsymbol{u}^c$ and $\boldsymbol{u}^m$ were 48×70 and 48×35, respectively. All inputs were normalized with L_2 regularization. The number of SOFA node pools was $5^6 = 15625$. The hidden state size of the node and raw clinical representations was 48×300. The spread breadth i was set to 3. We used Adam as the optimizer with a learning rate of 0.0005 and a batch size of 35. We also employed dropout on the outputs of the GGNN layer and RGRU layer with a dropout rate of 0.5. The number of epochs was 50.

4.3. Baseline methods

To validate the performance of the proposed DMLF model on mortality prediction, a comprehensive experiment was conducted with the following baseline models.

1) **SOFA Score**[2]: SOFA scores are widely used to determine the levels of organ dysfunction and mortality risk in ICU patients. We used tools from MDCalc to calculate the SOFA scores and mortality rates as one of our baselines. MDCalc is a collection of clinical decision tools and contents used by over 1.75 million medical professionals globally, including more than 65% of US physicians, every month [42].
2) **SAPS III**[3]: SAPS III is a system for predicting mortality and is one of several ICU scoring systems. It was designed to provide real-life predicted mortality rates for patients by following a well-defined procedure. Predicted mortalities are good when comparing groups of patients and possessing near-real-life mortalities. We also used tools from MDCalc to calculate the SAPS III scores and mortality rates.
3) **Machine Learning Baselines (MLBs)**. LR, a support vector machine (SVM), a random forest (RF), and XGBoost are traditional machine learning methods and are typically used as baselines in related works [3,4,17,43,44]. We use a toolkit from sklearn to reproduce our experiments with each of these methods.
4) **SAnD**: SAnD [17], a multitask model, employs a masked self-attention mechanism and uses positional encoding and dense interpolation strategies for mortality prediction.
5) **Multitask Channelwise LSTM (MC)**: MC is a long short-term memory (LSTM)-based multitask model that was proposed by Hrayr et al., 2019 [43]. MC has achieved the best mortality prediction performance in comparison with other baselines.
6) **Deep Interpretable Mortality Model (DIMM)**: The DIMM [3] is a multisource deep learning model explicitly designed for mortality prediction that utilizes GRU, multihead attention, and focal loss techniques.
7) **Gated Graph Sequence Neural Network (GGSNN)** The GGSNN [32] is the most recently proposed graph-based method and was designed to model sequential graphs. The GGSNN was used as one of the baselines to model SOFA graphs in this work.

4.4. Evaluation metrics

To evaluate the proposed framework comprehensively, we introduced tree evaluation metrics: 1) balanced accuracy (bAccuracy) [45]; 2) G-Mean [46]; 3) AUPRC.

4.5. Results

The comparative results are tabulated in Table 4. We grouped the results into four parts: the first part contains mortality prediction results from the ICU scoring systems, the second part includes mortality prediction results from the traditional data mining methods, the third part consists of the mortality prediction results from the explicitly designed cutting-edge deep learning methods, and the last part presents the results of our proposed method.

From the bAccuracy column, it is clear that most of the traditional data mining methods performed better than the scoring systems. Among the learning-based methods, deep learning methods outperformed the conventional data mining method by approximately 38.3%. This suggests that deep learning methods for mortality prediction can achieve better accuracy than clinical methods on the basis of daily reports.

SOFA had the lowest G-mean at approximately 58.08%, followed by the SAPS II score. This shows that G-mean is randomly attained when using scoring systems. LR, the SVM, and XGBoost were much better in terms of G-mean, achieving an average result of 63.68%. The deep learning methods achieved the best performance in terms of precision; this was especially true of our proposed framework, the DMLF, which exceeded 86% G-mean scores for all tasks.

For CCSID 2, the AUPRC scores followed a performance sequence of DMLF > GGSNN > DIMM > MC > LR > SVM > X GBoost > SAPS II > RF > SOFA > SAnD. This means that the scoring systems and the deep learning-based methods performed better than some data mining techniques based on their AUPRC scores. Undoubtedly, the proposed framework (the DMLF) achieved the best performance, which suggests that combining scoring systems and deep learning methods is effective. For other tasks, the performance was similar to that obtained on CCSID 2 on the basis of the recall rate.

From Table 3, we can see that the performance of different methods varied considerably across various tasks in terms of their AUROC values because the sample sizes and the weights of the learning features were different in different tasks, and most of these methods were not explicitly designed for mortality learning. Benefiting from the dual-core boosted mechanism, the DMLF could

[2] https://www.mdcalc.com/sequential-organ-failure-assessment-sofa-score.

[3] https://www.mdcalc.com/simplified-acute-physiology-score-saps-ii/.

Table 4
Performance comparison results of all baselines for mortality prediction.

CCS ID	CCS NAME	METHOD	bAccuracy	G-Mean	AUPRC
2	septicemia	SOFA	0.6310	0.5808	0.4791
		SAPS II	0.6761	0.6573	0.5706
		LR	0.6600	0.6185	0.6048
		SVM	0.6866	0.6503	0.6559
		RF	0.6524	0.6301	0.4582
		XGBoost	0.6784	0.6484	0.6000
		SAnD	0.6887	0.6465	0.4171
		MC	0.7904	0.7699	0.7801
		DIMM	0.8484	0.8420	0.8508
		GGSNN	0.8098	0.8184	0.7714
		DMLF	0.9185	0.8713	0.8611
49	diabetes	SOFA	0.6110	0.5024	0.3160
		SAPS II	0.6835	0.6450	0.4245
		LR	0.6007	0.4724	0.4867
		SVM	0.6152	0.4972	0.4744
		RF	0.6479	0.5887	0.3190
		XGBoost	0.6178	0.5041	0.4602
		SAnD	0.8885	0.8882	0.2474
		MC	0.6646	0.5932	0.5402
		DIMM	0.8534	0.8466	0.7122
		GGSNN	0.7155	0.6887	0.7643
		DMLF	0.8984	0.8953	0.8463
98	hypertension	SOFA	0.5810	0.4287	0.2507
		SAPS II	0.6680	0.6148	0.3837
		LR	0.5783	0.4140	0.4207
		SVM	0.5880	0.4331	0.4221
		RF	0.6166	0.5184	0.2879
		XGBoost	0.5795	0.4167	0.3790
		SAnD	0.7696	0.7601	0.2450
		MC	0.7802	0.7573	0.6159
		DIMM	0.8058	0.7877	0.7399
		GGSNN	0.8554	0.8471	0.7518
		DMLF	0.9107	0.8930	0.8323
101	atherosclerosis	SOFA	0.5892	0.4525	0.2379
		SAPS II	0.6724	0.6281	0.3508
		LR	0.5821	0.4190	0.4724
		SVM	0.5837	0.4197	0.4547
		RF	0.6355	0.5524	0.2925
		XGBoost	0.6259	0.5163	0.4432
		SAnD	0.6597	0.5688	0.2479
		MC	0.8496	0.8414	0.5775
		DIMM	0.8390	0.8275	0.7630
		GGSNN	0.6445	0.5434	0.7553
		DMLF	0.8727	0.8983	0.8523
157	renalfailure	SOFA	0.6297	0.5677	0.4413
		SAPS II	0.6806	0.6626	0.5250
		LR	0.6563	0.6017	0.5977
		SVM	0.6689	0.6075	0.6276
		RF	0.6539	0.6254	0.4400
		XGBoost	0.6808	0.6428	0.6105
		SAnD	0.7026	0.6655	0.4079
		MC	0.7626	0.7360	0.7732
		DIMM	0.6949	0.6406	0.8494
		GGSNN	0.8476	0.8443	0.7731
		DMLF	0.8897	0.8622	0.9023

mutually learn latent representations from various sources so that the impacts of different tasks could be dissolved to a great extent. We note that in our experiment, the running target was to achieve the best AUROC. Therefore, when compared with the DIMM, we did not obtain the best performance in the septicemia task in terms of the AUPRC. Overall, we achieved the best performance in all tasks among all baselines. This finding suggests that our method is the best approach and fits the mortality prediction task very well.

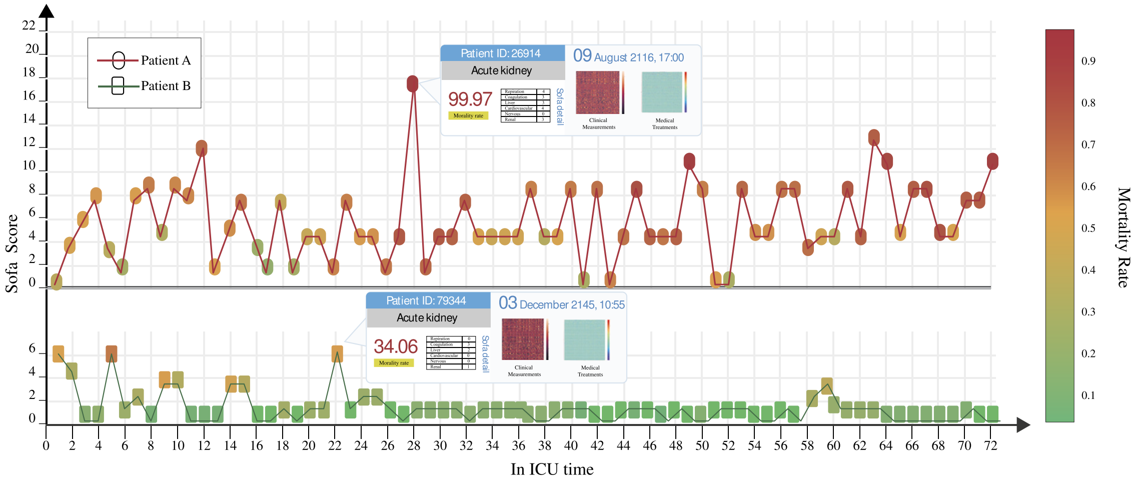


Fig. 3. Interpretability showcase between an SOFA scoring system and the DMLF mortality prediction method. Two patients were diagnosed with acute kidney disease; one died after 72 hours in the ICU (patient A), and the other was rescued after 72 hours in the ICU (patient B). The X-axis is the ICU stay time for patients, the y-axis is the SOFA score, and the color represents the mortality rate predicted by the DMLF.

4.6. Discussion and case study

The aim of this work was to bridge the gap between existing ICU scoring systems and cutting-edge deep learning methods in ICU mortality prediction problems. To actualize this goal, as listed in Section 1.1, we proposed two objectives. To assess our first goal, five evaluation metrics were adopted to evaluate our proposed method in comparison with existing ICU score systems. From Table 4, we find that our proposed method was 16.93% superior in terms of accuracy and 37.3% superior in terms of precision. Therefore, in the ICU mortality prediction scenario, the proposed method significantly outperformed the existing scoring system. Regarding method interpretability, as illustrated in Fig. 3, clinicians can evaluate and cross-check predicted mortality rates with an existing ICU scoring system utilized daily, such as SOFA. We can see from the figure that the DMLF prediction trend is overall in line with the SOFA scores, but the DMLF can provide more accurate prediction results. For example, for the same SOFA score of 4, the mortality rate predicted by DMLF for patient A was higher than that for patient B, and the patients' final results also confirmed this finding; i.e., patient A died, while patient B was rescued. Thus, we achieved the requirements of Objective 1.

To assess our second goal, we compared the prediction performance among the three groups of baselines. We can find that in terms of prediction performance, the machine learning methods were superior to ICU scoring systems. This indicates that incorporating machine learning into ICU mortality prediction can improve the prediction ability of the employed model and benefit the development of healthcare. Another finding is that the explicitly designed methods performed better than the general machine learning methods. The DMLF achieved the best performance, indicating that for higher mortality prediction performance, developing and customizing a suitable method is a necessity, and the simple reliance on generic methods cannot meet our requirements.

For the proposed method, the DMLF, we evaluated the two cores separately. For single-core modeling, the SOFA score contributed to the performance achieved on five different tasks and yielded an average accuracy of 0.815, but the precision varied greatly on the five different diseases. This indicates that the prediction process using only the SOFA score system is imprecise. For dual-core modeling, we achieved the best performance presented in Table 3. Moreover, the prediction performance was fairly stable (the variance was 4.45% vs. 16.93 for the state-of-the-art method). This finding indicates that our framework is on the cutting edge and is a reasonable paradigm. Furthermore, the SOFA scores can provide prediction evidence and reasonable explanations for the whole prediction process to support researchers, as these systems are very familiar to clinicians. Overall, we achieved Objective 2 proposed in Section 1.1. Despite the significant prediction performance and interpretability improvements achieved, our methods still have much room for further improvement. The first and the most critical concern remains the interpretability of the network. Although we enhanced its interpretability by using the SOFA scoring system, while this is a case-based explanation process, it did not improve the interpretability of the deep learning methods, so it cannot be integrated into our framework in an end-to-end manner. Another limitation is that our methods cannot be adopted for nonadults due to the insufficiency of the available clinical data.

5. Conclusion

In this paper, we proposed a unified dual-core mutual learning approach, the DMLF, that allows traditional ICU scoring systems to work together with raw EHRs to predict the mortality rates of ICU patients. In this work, we proposed two objectives: 1) to design a learning framework by utilizing existing scoring systems while preserving interpretability and 2) to achieve state-of-the-art prediction performance through mutual learning mechanisms. We used a graph neural network model to realize the first objective, and this model is regarded as one of the prediction scores of the proposed framework. To meet the second objective, we used an RGRU to learn directly from sequential raw EHRs, and this component is noted as another prediction score. We used a well-designed CSU and an attention mechanism to integrate these two scores into a unified learning framework to achieve the optimal prediction performance. Experimental results were obtained on a real-world dataset consisting of five tasks to verify the effectiveness and stability of the DMLF, and the findings demonstrated that our approach outperformed the baselines in terms of ICU mortality prediction.

All clinical decisions are based on evidence, we provide a visualization between ICU scoring systems' scores and the model's features weight, to the clinicians for our model to enhance the model's interpretability. 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 enhance the model's interpretability still remains a challenge, and our future work can focus on this problem.

CRediT authorship contribution statement

Zhenkun Shi: conceptualization, methodology, and writing original. **Sen Wang**: conceptualization and investigation. **Lin Yue**: writing – review & editing. **Yijia Zhang**: investigation and validation. **Binod Kumar Adhikari**: writing – review & editing. **Shuai Xue**: organized data, validate and analyze the results. **Wanli Zuo**: supervision. **Xue Li**: supervision. All authors analyzed the results and revised the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (no. 32201242), the China Postdoctoral Science Foundation (no. 2022M713328), the Tianjin Synthetic Biotechnology Innovation Capability Improvement Program (no. TSBICIP-CXRC-018) and the Nature Science Foundation of Jilin Province (no. 20190302029GX). The authors also gratefully acknowledge the financial support received from the China Scholarship Council (no. 201706170617).

References

[1] N. Nori, H. Kashima, K. Yamashita, H. Ikai, Y. Imanaka, 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, 2015, pp. 855–864.

[2] Z. Shi, W. Zuo, S. Liang, X. Zuo, L. Yue, X. Li, Iddsam: an integrated disease diagnosis and severity assessment model for intensive care units, IEEE Access (2020).

[3] Z. Shi, W. Chen, S. Liang, W. Zuo, L. Yue, S. Wang, Deep interpretable mortality model for intensive care unit risk prediction, in: International Conference on Advanced Data Mining and Applications, Springer, 2019, pp. 617–631.

[4] Z. Shi, W. Zuo, W. Chen, L. Yue, Y. Hao, S. Liang, Dmmam: deep multi-source multi-task attention model for intensive care unit diagnosis, in: International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 53–69.

[5] R. Joshi, P. Szolovits, Prognostic physiology: modeling patient severity in intensive care units using radial domain folding, in: AMIA Annual Symposium Proceedings, vol. 2012, American Medical Informatics Association, 2012, p. 1276.

[6] N. Liu, P. Lu, W. Zhang, J. Wang, Knowledge-aware deep dual networks for text-based mortality prediction, in: Proceedings - International Conference on Data Engineering, 2019, pp. 1406–1417.

[7] S. Ahmed, R.K. Chakrabortty, D.L. Essam, W. Ding, Poly-linear regression with augmented long short term memory neural network: predicting time series data, Inf. Sci. (2022).

[8] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, et al., The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults, Chest 100 (6) (1991) 1619–1636.

[9] M.M. Churpek, A. Snyder, X. Han, S. Sokol, N. Pettit, M.D. Howell, D.P. Edelson, Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit, Am. J. Respir. Crit. Care Med. 195 (7) (2017) 906–911.

[10] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (saps ii) based on a European/North American multicenter study, JAMA 270 (24) (1993) 2957–2963.

[11] H. Suresh, J.J. Gong, J.V. Guttag, Learning tasks for multitask learning: heterogeneous patient populations in the icu, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 802–810.

[12] X.S. Zhang, F. Tang, H.H. Dodge, J. Zhou, F. Wang, Metapred: meta-learning for clinical risk prediction with limited patient electronic health records, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2487–2495.

[13] Z. Shi, S. Wang, L. Yue, L. Pang, X. Zuo, W. Zuo, X. Li, Deep dynamic imputation of clinical time series for mortality prediction, Inf. Sci. 579 (2021) 607–622.

[14] D. Jiang, G. Tu, D. Jin, K. Wu, C. Liu, L. Zheng, T. Zhou, A hybrid intelligent model for acute hypotensive episode prediction with large-scale data, Inf. Sci. 546 (2021) 787–802.

[15] HCUP-US, Hcup clinical classifications software (ccs) for icd-9-cm, Website, https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp, 2021.

[16] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29, https://doi.org/10.1038/s41591-018-0316-z, http://www.ncbi.nlm.nih.gov/pubmed/30617335.

[17] H. Song, D. Rajan, J.J. Thiagarajan, A. Spanias, Attend and diagnose: clinical time series analysis using attention models, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[18] X. Liu, P. Hu, Z. Mao, P.-C. Kuo, P. Li, C. Liu, J. Hu, D. Li, D. Cao, R.G. Mark, et al., Interpretable machine learning model for early prediction of mortality in elderly patients with multiple organ dysfunction syndrome (mods): a multicenter retrospective study and cross validation, arXiv preprint, arXiv:2001.10977, 2020.

[19] Q. Xue, W. Zhang, H. Zha, Improving domain-adapted sentiment classification by deep adversarial mutual learning, arXiv preprint, arXiv:2002.00119, 2020.

[20] A.E.W. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.

[21] C.L. Chiang, The Life Table and Its Applications, Krieger, Malabar, FL, 1984.

[22] J. Xu, Y. Zhang, P. Zhang, A. Mahmood, Y. Li, S. Khatoon, Data mining on icu mortality prediction using early temporal data: a survey, Int. J. Inf. Technol. Decis. Mak. 16 (01) (2017) 117–159.

[23] R. Kuhn, O. Rahman, J. Menken, Survey measures of health: how well do self-reported and observed indicators measure health and predict mortality, in: Aging in Sub-Saharan Africa: Recommendations for Furthering Research, 2006, pp. 314–342.

[24] G.L. Sternbach, The Glasgow coma scale, J. Emerg. Med. 19 (1) (2000) 67–71.

[25] X. Zhang, B. Qian, X. Li, J. Wei, Y. Zheng, L. Song, Q. Zheng, An interpretable fast model for predicting the risk of heart failure, in: Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM, 2019, pp. 576–584.

[26] S. Yanamadala, D. Morrison, C. Curtin, K. McDonald, T. Hernandez-Boussard, Electronic health records and quality of care: an observational study modeling impact on mortality, readmissions, and complications, Medicine 95 (19) (2016).

[27] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE J. Biomed. Health Inform. 22 (5) (2017) 1589–1604.

[28] Z. Che, Y. Cheng, Z. Sun, Y. Liu, Exploiting convolutional neural network for risk prediction with medical feature embedding, arXiv preprint, arXiv:1701.07474, 2017.

[29] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor AI: predicting clinical events via recurrent neural networks, in: Machine Learning for Healthcare Conference, 2016.

[30] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085.

[31] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, IEEE J. Biomed. Health Inform. 22 (5) (2018) 1589–1604, https://doi.org/10.1109/JBHI.2017.2767063, https://www.ncbi.nlm.nih.gov/pubmed/29989977.

[32] D. Beck, G. Haffari, T. Cohn, Graph-to-sequence learning using gated graph neural networks, in: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), vol.1, 2018, pp. 273–283.

[33] J. Kim, T. Kim, S. Kim, C.D. Yoo, Edge-labeling graph neural network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11–20.

[34] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, D. Yin, Graph neural networks for social recommendation, in: The World Wide Web Conference, 2019, pp. 417–426.

[35] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, Others, scalable and accurate deep learning with electronic health records, npj Digit. Med. 1 (1) (2018) 18.

[36] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated Graph Sequence Neural Networks (1), 2015, pp. 1–20, arXiv:1511.05493, http://arxiv.org/abs/1511.05493.

[37] H. Luo, T. Li, B. Liu, J. Zhang, DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction, 2019, pp. 591–601, https://doi.org/10.18653/v1/p19-1056, arXiv:1906.01794.

[38] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, R. Jarina, Stacked convolutional and recurrent neural networks for music emotion recognition, arXiv preprint, arXiv:1706.02292, 2017.

[39] E. Choi, M.T. Bahadori, L. Song, W.F. Stewart, J. Sun, GRAM: graph-based attention model for healthcare representation learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 787–795.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[41] W. Wang, S.J. Pan, D. Dahlmeier, X. Xiao, Coupled multi-layer attentions for co-extraction of aspect and opinion terms, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[42] MD+CALC, About mdcalc, Website, https://www.mdcalc.com/about-us, 2020.

[43] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 1–18.

[44] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[45] H. Carrillo, K.H. Brodersen, J.A. Castellanos, Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy, in: ROBOT2013: First Iberian Robotics Conference: Advances in Robotics, vol. 1, Springer, 2014, pp. 347–361.

[46] R.P. Espíndola, N.F. Ebecken, On Extending f-Measure and g-Mean Metrics to Multi-Class Problems, WIT Transactions on Information and Communication Technologies, vol. 35, 2005, pp. 25–34.