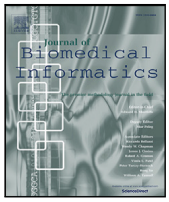
Journal of Biomedical Informatics 143 (2023) 104408

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Original Research

LGTRL-DE: Local and Global Temporal Representation Learning with Demographic Embedding for in-hospital mortality prediction

Mengjie Zou [a], Ying An [b], Hulin Kuang [a], Jianxin Wang [a,*]

[a] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, PR China
[b] The Institute of Big Data, Central South University, Changsha, 410083, PR China

ARTICLE INFO

Keywords:
Deep learning
Electronic medical records
In-hospital mortality prediction
Local and global temporal representation
Demographic embedding

ABSTRACT

Predicting the patient's in-hospital mortality from the historical Electronic Medical Records (EMRs) can assist physicians to make clinical decisions and assign medical resources. In recent years, researchers proposed many deep learning methods to predict in-hospital mortality by learning patient representations. However, most of these methods fail to comprehensively learn the temporal representations and do not sufficiently mine the contextual knowledge of demographic information. We propose a novel end-to-end approach based on Local and Global Temporal Representation Learning with Demographic Embedding (LGTRL-DE) to address the current issues for in-hospital mortality prediction. LGTRL-DE is enabled by (1) a local temporal representation learning module that captures the temporal information and analyzes the health status from a local perspective through a recurrent neural network with the demographic initialization and the local attention mechanism; (2) a Transformer-based global temporal representation learning module that extracts the interaction dependencies among clinical events; (3) a multi-view representation fusion module that fuses temporal and static information and generates the final patient's health representations. We evaluate our proposed LGTRL-DE on two public real-world clinical datasets (MIMIC-III and e-ICU). Experimental results show that LGTRL-DE achieves area under receiver operating characteristic curve of 0.8685 and 0.8733 on the MIMIC-III and e-ICU datasets, respectively, outperforming several state-of-the-art approaches.

1. Introduction

Healthcare is one of the remarkable applications of data mining and machine learning, which is closely related to people's health and welfare, and has been highly concerned by the government. Critically ill patients receive cares in the Intensive Care Unit (ICU) with advanced diagnostic and therapeutic techniques, and require close monitoring. Early assessment of patients' health status and predicting in-hospital mortality can assist doctors to make clinical decisions, improve patients' outcomes and allocate clinical resources effectively.

The widespread adoption of rich digital clinical systems, primarily Electronic Medical Records (EMRs), which store large amounts of data, not only improves hospital management and services but also facilitates the advancement of data-driven approaches. As shown in Fig. 1, EMR data is a collection of time-stamped clinical event tables that store the longitudinal medical data closely related to the patient, including laboratory tests, vital signs, etc. Besides, EMR contains patients' demographic information such as age and gender. Due to the comprehensiveness and accessibility of EMRs, it has become a hot spot in medical data analysis and has been widely used in various clinical tasks such as in-hospital mortality prediction [1–3], hospital length-of-stay prediction [4,5], readmission prediction [6,7], and disease risk prediction [8–10]. Among them, in-hospital mortality prediction is closely related to intervention selection, care planning, and medical resource allocation. Accurate evaluation of mortality is the key to improving survival rate and physiological outcomes of patients.

Most of the early prediction methods for in-hospital mortality based on machine learning rely on traditional human-intervention feature engineering, resulting in insufficient scalability and generalization [11]. Therefore, deep learning methods with excellent automatic feature learning capability have gradually become mainstream in the field of medical data analysis and applications. In recent years, many scholars regard longitudinal EMRs as multivariate time series and use deep learning methods to extract a relevant representation of the patient health status for downstream clinical prediction tasks [12–14]. Although these methods show promising application prospects, they still have some shortcomings in capturing the deep dependencies between clinical events. More specifically, there are two main challenges as:

* Corresponding author.
E-mail addresses: mengjyzou@csu.edu.cn (M. Zou), anying@csu.edu.cn (Y. An), hulinkuang@csu.edu.cn (H. Kuang), jxwang@mail.csu.edu.cn (J. Wang).

https://doi.org/10.1016/j.jbi.2023.104408
Received 24 May 2022; Received in revised form 28 March 2023; Accepted 28 May 2023
Available online 7 June 2023

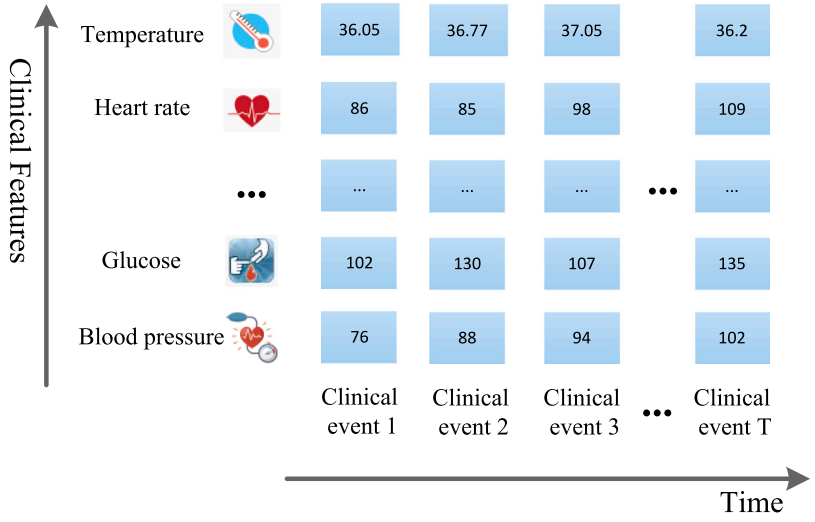


Fig. 1. The example of a patient's electronic medical record.

C1. Capturing temporal dependencies of EMR data

EMR data of a single patient often contains multiple clinical events, which are not isolated sets but are dependent on each other. In clinical practice, doctors usually diagnose patients according to their historical clinical events, and then take corresponding treatment measures. In addition, abnormalities in a patient's clinical events at one time step may be related to previous clinical events and have lasting effects on subsequent clinical events. For example, the elevated body temperature of patients may be closely related to their previous viral infection, and may subsequently lead to clinical events such as elevated heart rate and blood sugar. The dependencies among clinical events reflect the development pattern of the patient health status, so effectively capturing these dependencies is critical to achieving accurate in-hospital mortality prediction.

Generally, the dependencies among clinical events can be divided into two categories: local dependencies and global dependencies. The local dependencies refer to the correlations between adjacent clinical events. They can help us learn the short-term variations of patient health status, which is important for critically ill patients because their health status is always in change. The global dependencies refer to the relationships between each clinical event and the rest of the clinical events during a complete hospitalization, which can be used to assess the patient's health status in a global view by integrating the outcomes of all clinical events.

Most of the existing studies are based on Recurrent Neural Network (RNN) architectures such as Long Short-Term Memory (LSTM) network and Gated Recurrent Unit (GRU) to learn the temporal dependencies of EMR data [15]. Despite their successful application on many clinical tasks, there are two common problems, namely gradient vanishing and exploding problems, which limit their ability to learn long-term dependencies [16]. In the latest research, parallel Convolutional Neural Network (CNN) and attention mechanisms are also transferred to learn dependencies among clinical events across time scales and achieve early success in model training efficiency and context information extraction [17]. Although they alleviate the long-term dependence problem, CNNs focus more on learning the local information of EMR due to their convolutional operational characteristics and are difficult to directly compute the dependencies between clinical events. More recently, Transformer has been proposed for natural language processing tasks and shows considerable sequence learning capabilities [18]. Since each input vector in Transformer yields a new vector containing information of the original vector and other vectors, Transformer is used by some clinical prediction tasks to extract the global dependencies of EMR data [19–21]. The above methods usually consider only one aspect of temporal dependencies. In order to effectively capture temporal relationships of EMR data, it is necessary to design a model that considers both local and global dependencies among clinical events.

C2. Extracting background information of the patient health status

The patient's demographic data is often considered to be non-temporal static background information and used to assess the patient's basic condition. For the same vital sign, different demographic information reflects different conditions. For example, the same blood pressure value that is normal for a young patient may be high for an older patient [22,23]. Therefore, it is important to take full account of patient demographic information when making clinical predictions. However, the demographic data has either been ignored [17,24] or simply concatenated together with other clinical time-series features [8, 15] in existing studies. They fail to sufficiently mine the contextual knowledge in static information. Therefore, it is necessary to design a more adaptive approach to fully consider the impact of demographic information.

To address the above challenges, in this study, we propose a novel approach based on Local and Global Temporal Representation Learning with Demographic Embedding, called LGTRL-DE, for in-hospital mortality prediction. To solve the challenge C1, we first utilize a local temporal representation learning module combining Bidirectional Gate Recurrent Unit (Bi-GRU) and Local Attention Mechanism (LAM) to learn the short-term local features of patients' health status. Besides, we design a Transformer-based global temporal representation learning module to further capture the long-term global dependencies among different clinical events, thus obtaining a more comprehensive semantic representation for the patient. To solve the challenge C2, we take the patient's demographics as the background information of health status and use it to initialize the Bi-GRU hidden state units in the local temporal representation learning module to deeply mine the correlations between patient condition progression and its demographic features. Meanwhile, a multi-view representation aggregation module is proposed to fuse the demographic features with the local and global temporal dependencies to further improve the effectiveness of the final patient representations.

Our main contributions are summarized as follows:

- We propose a novel end-to-end patient health status representation learning framework called LGTRL-DE, which utilizes three modules to fully extract the temporal relationships of EMR data and learns comprehensive representations for in-hospital mortality prediction.

- We design a local temporal representation learning module by combining bidirectional GRU and local attention mechanism, and further use the Transformer-based global temporal representation module to learn the global dependencies among clinical events.
- We take the demographic data as contextual information to initialize the hidden units in the local temporal representation learning module, and then fully integrate the dynamic temporal features and static demographic features to obtain a personalized patient health status representation.

2. Related work

2.1. Temporal feature extraction for EMR data

EMR can usually be viewed as a time-series organized in chronological order by multiple historical clinical events of the patient. In recent years, deep learning methods have been widely used in the analysis and application of EMR data. A large number of RNN- and CNN-based methods have been proposed to mine the temporal patterns of disease progression from patient clinical records [4,17,24–28]. For example, Retain [26] proposed to simulate doctor's behavior in clinical practice by using two LSTMs to summarize the EMR data in forward and reverse order chronologically, respectively. ConCare [25] used multi-channel GRU to capture the correlation between dynamic features and static features, and designed a time-aware mechanism to learn the information decay from irregular time intervals. AdaCare [17] proposed a multi-scale adaptive recalibration module with dilated convolution to capture biomarker features of clinical time-series. StageNet [24] proposed a stage-aware LSTM module and a stage-adaptive convolutional module to extract disease stage variations and incorporate them into risk prediction respectively. TPC [4] proposed a new deep learning model based on the combination of temporal convolution and point-wise convolution to demonstrate its better performance in clinical prediction tasks. Most of the above methods are based on CNN and RNN models, which cannot effectively capture the global information of clinical events due to the long-term dependence problems of RNN and the convolutional operational characteristics of CNN. Thus, in this study, we propose a new representation learning method to capture both local and global dependencies between different clinical events from patient EMR data, and deeply incorporate the static health context information hidden in patient demographic data to further improve the effectiveness and comprehensiveness of patient health status representations.

2.2. Attention mechanism for healthcare

In clinical practice, clinical indicators contribute differently to clinical outcomes, and traditional temporal models cannot fully capture important information in EMR. In recent years, attention mechanisms have achieved initial success in fields such as natural language process and computer vision, and gradually transferred to healthcare applications [15]. Attention-based models learn a weight for each event, which enables the models to effectively focus on events that are more important to patient disease development. Dipole [29] used Bidirectional Recurrent Network (Bi-RNN) and three attention mechanisms to learn patient representations from historical visits. The three attention mechanisms were introduced to measure the relationships of different visits for prediction. Unlike the transitional attention mechanism, Vaswani et al. [18] proposed the self-attention mechanism to capture the relationship between words at arbitrary distances for the machine translation task. By applying the self-attention mechanism, each input vector can get a new vector containing information from the original vector and other vectors, which enables the model to consider the global information and solves the deficiency of the traditional attention mechanism. Inspired by [18], Song et al. [19] presented a self-attention-based architecture named SAnd for clinical time-series modeling and achieved comparable performance to the LSTM model. STraTS [21] was also a self-attention-based architecture for healthcare that solved the data missing by avoiding aggregation and imputation. In addition, INPREM [30] used a two-level attention mechanism to extract correlations between visits and provided interpretability through an attention map generated by the self-attention mechanism. In our work, we also use the self-attention mechanism to learn correlations between clinical events and enhance the health status representations of patients.

3. Methodology

3.1. Basic notations and problem definition

3.1.1. Notations

To clarify our approach more clearly, we summarize the notations used in this paper. In the longitudinal EMR data, each patient can be represented as a sequence of multivariate clinical time-series $P = [r_1, r_2, \dots, r_t, \dots, r_T]$, where r_t is the tth clinical event and T is the number of clinical events. Each clinical event $r_t = [v_1, v_2, \dots, v_f, \dots, v_F] \in \mathbb{R}^F$ consists of a bag of clinical features (e.g., vital signs and laboratory measurements, etc.) including numerical and categorical variables, where v_f represents the fth clinical feature and F is the total number of clinical features. For all the clinical features, some are continuous variables, and some are discrete variables. We treat discrete variables as one-hot vectors and concatenate them with continuous variables. In addition, the non-temporal demographic information is denoted as $S = [s_1, s_2, \dots, s_m, \dots, s_M] \in \mathbb{R}^M$, where s_m is the mth demographic feature (e.g., age and gender, etc.) and M is the number of demographic features.

3.1.2. Problem definition

Given a patient's historical clinical time-series $P \in \mathbb{R}^{T \times F}$ and demographic information $S \in \mathbb{R}^M$ based on the first 48 h of an ICU stay, our goal is to learn a health status representation from P and S, and then use this representation to predict whether the patient has a risk of mortality during this ICU stay. This problem is defined as a binary classification task, denoted as $y = LGTRL\text{-}DE(P, S) \in \{0, 1\}$.

3.2. Model overview

As shown in Fig. 2, our proposed LGTRL-DE is an end-to-end deep learning network, which consists of the following three modules: (1) The local temporal representation learning module (LTRLM) is used to extract the local temporal dependencies between different clinical events from patient's historical clinical records. (2) The global temporal representation learning module (GTRLM) is developed to acquire global information about the patient that is combined with the local temporal dependencies to enhance the patient's representation. (3) The multi-view representation fusion module (MRFM) is adopted to summarize the representation of all time steps by a gate control mechanism, and fuse static demographic information to achieve personalized in-hospital mortality prediction.

3.3. Local Temporal Representation Learning Module (LTRLM)

In this module, we extract local dependencies from the embedded clinical time-series input by combining Bi-GRU using the projected demographic information as initialization and the proposed local attention mechanism module.

3.3.1. Clinical time-series embedding

Given the input matrix P, we first embed each clinical event $r_t \in \mathbb{R}^F$ into a continuous, dense vector representation $x_t \in \mathbb{R}^n$ through a 1-layer feed-forward network (FFN), which can be denoted as $x_t = ReLU(W_x r_t + b_x)$, where $W_x \in \mathbb{R}^{n \times F}$ is the weight matrix of clinical event in medical records, $b_x \in \mathbb{R}^n$ is the bias vector, and n is the size of embedding dimension. Thus, the original input P is mapped into a clinical embedding vector sequence $[x_1, x_2, \dots, x_t, \dots, x_T]$.

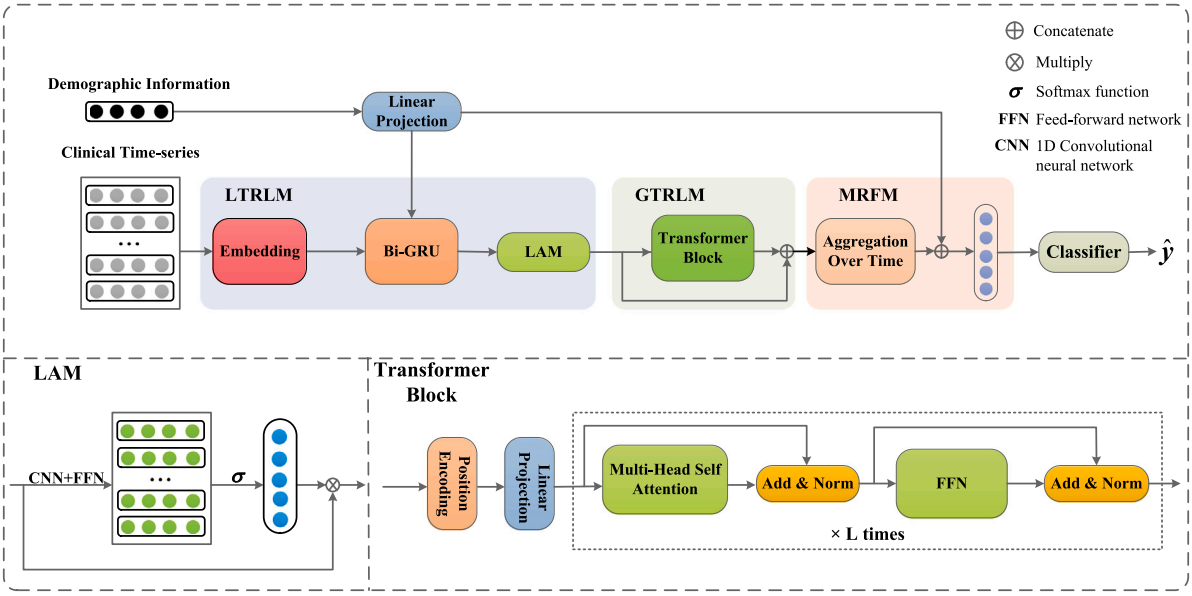


Fig. 2. An overview of the proposed LGTRL-DE. LTRLM represents the local temporal representation learning module. GTRLM represents the global temporal representation learning module. MRFM represents the multi-view representation fusion module. Firstly, the clinical time-series and demographic data are used to generate a patient context matrix through the local temporal representation learning module (LTRLM). Then, we further learn the global dependencies among clinical events through the global temporal representation learning module (GTRLM). MRFM is used to further aggregate the representations of all time steps and fuse the demographic data again to obtain the final patient representation. Finally, the patient representation is fed into the Classifier to obtain the prediction result.

3.3.2. Demographic information embedding

As described before, a patient's demographic information can be seen as important background data that indicates the patient's basic health condition. In order to effectively combine the static demographic information and dynamic clinical time series, we first encode the demographic information $S \in \mathbb{R}^M$ into the same space as the clinical embedding x_t by linear projection. It can be denoted as $d = W_d S$, where $d \in \mathbb{R}^n$ is the embedding vector of demographic information, $W_d \in \mathbb{R}^{n \times M}$ is learnable projection matrix.

3.3.3. Bi-GRU with demographic initialization for temporal learning

We use a Bi-RNN to capture temporal relationships of EMR data from the clinical embedding vector sequence. RNNs have several variants such as LSTM [31] and GRU [32]. Similarly, Bi-RNN has two main variants, Bidirectional Long Short-Term Memory (Bi-LSTM) and Bi-GRU. In our implementation, we find that the Bi-GRU shows better performance than Bi-LSTM and has a simpler structure, so we employ Bi-GRU to capture clinical temporal dependencies.

The traditional RNN model has the hidden state cell initialized to zero or random by default, which fails to incorporate the contextual information. To solve this problem, we use demographic embedding to initialize the hidden state cell of Bi-GRU to enhance semantic information. Taking the forward GRU as an example,

$$\overrightarrow{h}_0 = d \tag{1}$$

where $\overrightarrow{h}_0$ is the initial forward hidden state. Next, for clinical embedding representation x_t and its previous hidden state $\overrightarrow{h}_{t-1}$, we compute the forward hidden state $\overrightarrow{h}_t$ at time t as:

$$\overrightarrow{h}_t = GRU(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

In a similar way, we can also obtain the hidden state $\overleftarrow{h}_t$ of a backward GRU cell. The number of hidden units for both forward and backward GRUs is $n/2$. Then, the final hidden state h_t of a Bi-GRU cell can be obtained by concatenating the forward hidden state $\overrightarrow{h}_t$ and the backward hidden state $\overleftarrow{h}_t$ as:

$$h_t = Concat(\overrightarrow{h}_t, \overleftarrow{h}_t) \tag{3}$$

where $h_t \in \mathbb{R}^n$ represents the tth hidden state of Bi-GRU and $Concat$ denotes the concatenation operation.

3.3.4. Local attention mechanism

In clinical practice, different clinical events may have different effects on the development of patients' future health status. To simulate this situation in the model, we employ a Local Attention Mechanism (LAM) block to learn the context representations which contain the importance of the patient's clinical events. Concretely, in our local attention mechanism, since the input is a feature vector containing the hidden states of all clinical events (i.e., all time steps), we first perform a 1D Convolutional layer with a kernel size of 3 and a stride of 1 to learn the relationship between each clinical event and its two adjacent clinical events (the previous and next clinical events). In this CNN layer, the number of filters is 128, which is decided experimentally. Then, to adaptively average the learned relationship of the 3 adjacent clinical events, we perform a 1-layer feed-forward network (FFN) on the output of the convolution layer. Finally, to compute the attention weight map, we use a Softmax function on the output of FFN. The above process can be formulated as:

$$a_t = \sigma(W_a^T(\mathbb{C}^K \cdot h_t) + b_a) \tag{4}$$

where $\mathbb{C}^K$ is the 1D convolution operation with a kernel size of K (K = 3 in this study), $a_t \in \mathbb{R}$ represents the local attention weight of the corresponding clinical event, σ represents Softmax function, $W_a \in \mathbb{R}^n$ and $b_a \in \mathbb{R}$ are the learned weights and biases, respectively.

Finally, the hidden state h_t is multiplied by the local attention weight a_t to generate the context vector representation c_t:

$$c_t = a_t \cdot h_t \tag{5}$$

In general, the local information is different among different time events, so that a_t differs and each dimension in c_t becomes different from that in h_t, leading to changes in the correlation structure between h_ts. Besides, the computed c_t also carries the information of the local dependency rather than only the original correlation structure between h_ts, because a_t has captured the local information between three adjacent events. As a result, the original clinical time-series of the patient can be represented by a context matrix $C = [c_1, c_2, \ldots, c_t, \ldots, c_T] \in \mathbb{R}^{T \times n}$.

3.4. Global Temporal Representation Learning Module (GTRLM)

After obtaining the local dependencies among clinical events through LTRLM, we utilize a Transformer-based global temporal representation learning module to further learn the global dependencies among clinical events, and combine local and global features to enhance the patient representation.

Transformer allows each clinical event to focus on other clinical events directly and extracts the relevant contextual information from the other clinical events regardless of their distance in the sequence. Fig. 2 shows the architecture of the Transformer block. It mainly contains the position encoding and L successive layers with a Multi-Head Self-Attention (MHSA) followed by a two-layer feed-forward network. Since clinical events in a medical record are sequential and Transformer is insensitive to the ordering of input, we first add positional encoding to the context vector representations:

$$e_t = c_t + p_t \tag{6}$$

where e_t is the encoded context representation vector, and p_t is the positional encoding vector, $p_{t,2k-1} = cos(t/10000^{\frac{2k}{n}})$ and $p_{t,2k} = sin(t/10000^{\frac{2k}{n}})$ for $1 \leqslant 2k-1 < 2k \leqslant n$. Before being fed into the MHSA block, the encoded context representations $e_t \in \mathbb{R}^n$ are packed into the matrix $E \in \mathbb{R}^{T\times n}$, which will be input into the linear projection layer to generate the initial matrix Φ_0, where the subscript 0 denotes the first layer of Transformer.

Then, we employ the MHSA with h heads to capture the hidden dependencies between clinical events. Concretely, we put Φ_0 into the following equation:

$$\widetilde{\Phi}_l = MHSA(Norm(\Phi_{l-1})) + \Phi_{l-1} \tag{7}$$

where $1 \leqslant l \leqslant L$ and Φ_{l-1} denotes the output of the $(l-1)$th Transformer layer, and $Norm$ represents the normalization layer [33].

After that, we employ a feed-forward network to transform the representation non-linearly as follows:

$$\Phi_l = FFN(Norm(\widetilde{\Phi}_l)) + \widetilde{\Phi}_l \tag{8}$$

where FFN consists of two fully-connected layers and is defined as $FFN(x) = W_2(ReLU(W_1 x + b_1) + b_2)$. To prevent over-fitting, we add a dropout layer [34] behind the FFN. Since we stack the Transformer layer L times, we use the results obtained in the last layer as the final clinical events' global representations, which can be denoted as Φ_L.

Finally, we concatenate the context representation matrix C and the global representation matrix Φ_L to obtain the enhanced patient representations:

$$U = Concat(C, \Phi_L) \tag{9}$$

As a result, the output of this module is a enhanced matrix $U = [u_1, u_2, \ldots, u_t, \ldots, u_T] \in \mathbb{R}^{T\times 2n}$ and u_t is the enhanced representation at the tth time step.

3.5. Multi-view Representation Fusion Module (MRFM)

In this module, we generate a dense and unified representation that fuses temporal clinical events and static demographic information to perform the final personalized prediction. Firstly, we use Bi-GRU to aggregate the vectors of all time steps from forward and backward to obtain dense semantic information g_T from clinical events:

$$g_T = Bi\text{-}GRU(u_1, u_2, \ldots, u_t, \ldots, u_T) \tag{10}$$

where Bi-GRU has $n/2$ hidden state units and $g_T \in \mathbb{R}^n$ denotes the final dense semantic representation. The unique gate control mechanism in the GRU cell can effectively aggregate the vector of the multiple time steps by forgetting the useless information and keeping important features. Then, to incorporate the patient's demographic information, we fuse the dense semantic representation g_T and the embedding vector of demographic information d together by concatenation operation as:

$$z = Cancat(g_T, d) \tag{11}$$

where $z \in \mathbb{R}^{2n}$ is the final patient representation.

3.6. Classification and loss function

Finally, we feed z into a fully-connected layer with $sigmoid$ function to predict the final mortality $\hat{y}$ as follows:

$$\hat{y} = sigmoid(W_y z + b_y) \tag{12}$$

where W_y and b_y are the learnable parameters of the fully-connected layer. Since our task is a binary classification problem, we employ binary cross-entropy to calculate the loss of the prediction $\hat{y}_i$ and ground-truth label y_i of each patient as follow:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{13}$$

where θ represents all trainable parameters in the model.

4. Experiments

In this section, we evaluate the performance of our proposed model by comparing it with several state-of-the-art methods.

4.1. Dataset description and data processing

In this study, we use the following two datasets for the in-hospital mortality prediction task, and Table 1 shows the data description.

MIMIC-III Dataset. The Medical Information Mart for Intensive Care III (MIMIC-III[1]) is a public intensive care database maintained by the Massachusetts Institute of Technology (MIT) Laboratory of Computational Physiology [35]. It includes electronic medical records of more than 40,000 patients admitted to the Intensive Care Unit (ICU) between 2001 and 2012.

We process the raw data according to the previous work [36]. The target of the task is to predict the in-hospital mortality based on the first 48 h of an ICU stay. We define this task as a binary classification and the ground-truth label indicates whether the patient died before discharge from the ICU. We select all the adult patients (not less than age 18) with an ICU stay longer than 48 h and obtain a cohort including 21,139 unique ICU stays in our final dataset, which contains 18,342 negative samples (alive) and 2797 positive samples (death). The reason of why we use the first 48 h are as follows: (1) The clinical demand of making the decision and allocating clinical resources in advance requires using a few first hours rather than the whole ICU stay; (2) Using too many or too few hours may decrease the performance [14]; (3) Previous studies have demonstrated that using the first 48 h is sufficient and more effective than using fewer hours [36,37].

e-ICU Dataset. The eICU Collaborative Research database (e-ICU[2]) is a multi-center intensive care database that contains more than 200,000 ICU stays collected between 2014 and 2015 from 208 hospitals located in the United States [38] .

Similar to the processing of the MIMIC-III dataset, we use the same cohort selection criteria. After filtering and data cleaning, a cohort containing 30,680 unique ICU stays is reserved as the final dataset. The distribution of positive and negative samples is shown in Table 1.

We conduct the prediction based on the first 48 h of an ICU stay and each hour is considered as a time step. For each time step, there is a clinical event containing several variables. If a variable in the clinical event of a time step has multiple measurements, we use the value

[1] https://physionet.org/content/mimiciii/1.4/
[2] https://physionet.org/content/eicu-crd/2.0/

Table 1
Cohort summaries of MIMIC-III and e-ICU datasets.

	MIMIC-III dataset			e-ICU dataset		
	Negative	Positive	Total	Negative	Positive	Total
The number of ICU stays	18,342	2,797	21,139	27,157	3,523	30,680
Gender(male %)	56.3%	53.8%	56.0%	55.0%	52.5%	54.7%
Age(mean)	63.81	64.48	63.90	62.21	63.95	62.41
Number of temporal clinical features	17			20		
Number of demographic features	6			6		

of the last measurement. Besides, followed by [36,39], we perform imputation to deal with the missing values in both datasets as follows. For clinical features, we impute the missing values using the value at the previous time-step if they exist and otherwise impute the missing values using the average values over all selected patients in the dataset. For demographic features, we also impute the missing values using the average values over all selected patients. For clinical features, according to the feature selection methods in [36,39], we select 17 and 20 clinical variables on the MIMIC-III and e-ICU datasets, respectively. For demographic features, on both datasets, we select 6 demographic features: age, gender, height, weight, ethnicity, and admission type which have been demonstrated to be important for in-hospital mortality prediction in previous studies [25,27]. The descriptive statistics of the clinical features and demographic information in the MIMIC-III and e-ICU datasets are shown in Supplementary Material.

4.2. Baseline models

We compare the proposed LGTRL-DE with the following baseline models.

Bi-GRU. It is a standard bidirectional gated recurrent unit network and we use the last time step's hidden state vector to execute the mortality prediction.

$Transformer_e$ [18]. It is the encoder of the Transformer. We flatten the last encoder's output and use an FFN to make the prediction.

Retain [26]. Retain is based on a two-level neural attention model that detects influential clinical events and important clinical variables.

SAnd [19]. SAnd is a masked self-attention-based architecture for clinical time-series analysis and employs the dense interpolation strategy for incorporating temporal information.

STraTS [21]. STraTS is a Transformer-based self-supervised architecture for clinical time-series learning. We modify the last layer of the model so that it is adapted to our supervised learning task.

INPREM [30]. INPREM is designed with a feed-forward network-based linear module and a Transformer-based nonlinear module that provide interpretability and dependency modeling, respectively.

ConCare [25]. ConCare embeds the feature sequences separately by multi-channel GRUs and uses the multi-head self-attention to capture the inter-dependencies among dynamic features and static baseline information.

AdaCare [17]. AdaCare utilizes the dilated convolutions with multi-scale receptive fields and gated recurrent units to capture the long and short-term temporal information of biomarkers.

Since most of the original baseline models do not consider the demographic information. For a fair comparison of performance, we add the embedding vector d of demographic information into the final hidden representation of these models for mortality prediction.

4.3. Implementation details

Our proposed LGTRL-DE is implemented in Tensorflow framework with Python 3.7.0 in NVIDIA RTX 2080Ti GPU. The parameters are trained by Adam optimizer with a learning rate of 5×10^{-4} and the mini-batch is set to 256. The dropout strategy is employed to avoid over-fitting and the dropout rate is set to 0.5. The size of embedding dimension n is set as 128 after extensive experiments using different sizes. The number of Transformer layers $L = 4$ and the number of heads h in each multi-head self-attention is 2. The value of T is 48 for both datasets. We train each model using 40 epochs on the MIMIC-III dataset and 100 epochs on the e-ICU dataset, repsectively. To better evaluate the performance of the model, we conduct 5-fold cross-validation on both two datasets. For a fair comparison of performance, the implementation details of other baselines follow their instructions, and the optimal hyperparameters are tuned on the same training and validation sets. The source code is publicly available at GitHub.[3]

4.4. Evaluation metrics

Since the in-hospital mortality prediction is defined as a binary classification task and the datasets are imbalanced, we use the area under receiver operating characteristic curve (AUROC), the area under precision–recall curve (AUPRC), min(Se, P+) which is the minimum of precision and sensitivity, and F1-score to evaluate the performance of methods in our experiments. Among them, the min(Se, P+) is the custom metric presented in the Physionet/CinC Challenge 2012 [40]. F1-score is the reconciled mean of recall and precision. AUROC and AUPRC are the most informative evaluation metrics for imbalanced data classification. The larger their values, the stronger the ability of the model to distinguish positive and negative samples. Besides, we save the best model which achieves the largest AUROC value on the validation set during model training and then evaluate it on the test set to report its performance. We report the mean and standard deviation of the performance measures on the test set for both two datasets.

5. Results

5.1. Comparison with baselines

Table 2 reports the performances of the proposed LGRTL-DE and all baseline models on MIMIC-III and e-ICU datasets. We can observe that our proposed model LGRTL-DE achieves the best performance among all compared methods on both MIMIC-III and e-ICU datasets. In particular, the AUPRC values of our model on the two datasets reach 0.5607 and 0.5801, which improves the best baseline by 1.83% and 2.13%, respectively.

From Table 2, we can see that Bi-GRU and $Transformer_e$ methods obtain relatively good performance on both datasets, and the value of AUROC are both 0.8582 on the MIMIC-III dataset. The result of $Transformer_e$ further illustrates its reasonableness and effectiveness for clinical time-series learning. Due to the complexity of EMR data, Retain only relying on two unidirectional RNNs is difficult to capture the dependencies between clinical events, and thus the performance is worse than the basic Bi-GRU. Besides, SAnd utilizes the dense interpolation method to obtain a unified representation of the sequence, but it fails to fully capture the temporal dependencies of clinical events, which results in its relatively low F1-score among all baselines. Like SAnd, STraTS uses a self-attention layer to compute the time-series embedding by directly summarizing after deriving the weights, which ignores the sequential information of the clinical sequence and has the lowest

[3] https://github.com/Mengjielf/LGTRL-DE

Table 2
Performance comparisons with eight baselines on the MIMIC-III and e-ICU datasets. The number in () indicates the standard deviation.

Methods	MIMIC-III dataset				e-ICU dataset			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
Bi-GRU	0.8582(.006)	0.5418(.015)	0.5201(.016)	0.4559(.020)	0.8665(.006)	0.5588(.018)	0.5417(.019)	0.4799(.030)
Retain [26]	0.8455(.006)	0.5013(.016)	0.4962(.008)	0.4311(.027)	0.8651(.006)	0.5586(.016)	0.5362(.016)	0.4596(.034)
$Transformer_e$ [18]	0.8582(.004)	0.5375(.018)	0.5222(.015)	0.4381(.025)	0.8629(.005)	0.5554(.015)	0.5420(.020)	0.4928(.018)
SAnd [19]	0.8566(.002)	0.5360(.020)	0.5178(.016)	0.4098(.054)	0.8626(.003)	0.5492(.018)	0.5325(.026)	0.4729(.043)
STraTs [21]	0.8444(.006)	0.5030(.017)	0.4937(.009)	0.3964(.053)	0.8599(.003)	0.5392(.020)	0.5234(.017)	0.4773(.019)
INPREM [30]	0.8557(.004)	0.5265(.032)	0.5172(.012)	0.4280(.053)	0.8544(.005)	0.5235(.022)	0.5164(.021)	0.4534(.039)
ConCare [25]	0.8632(.008)	0.5342(.006)	0.5174(.025)	0.4473(.047)	0.8617(.009)	0.5443(.017)	0.5282(.023)	0.4086(.024)
AdaCare [17]	0.8608(.006)	0.5424(.012)	0.5205(.014)	0.4716(.027)	0.8678(.008)	0.5568(.012)	0.5398(.010)	0.4832(.023)
LGTRL-DE	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

Table 3
Performance comparison for LGTRL-DE's variants. The number in () indicates the standard deviation.

Methods	Modules			MIMIC-III dataset				e-ICU dataset			
	LTRLM	GTRLM	MRFM	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
$LGTRL\text{-}DE_{L-}$	✗	✓	✓	0.8525(.006)	0.5231(.019)	0.5126(.010)	0.4093(.041)	0.8675(.006)	0.5639(.018)	0.5400(.020)	0.5037(.018)
$LGTRL\text{-}DE_{G-}$	✓	✗	✓	0.8605(.003)	0.5452(.006)	0.5185(.012)	0.4651(.010)	0.8696(.005)	0.5691(.016)	0.5493(.022)	0.4889(.035)
$LGTRL\text{-}DE_{M-}$	✓	✓	✗	0.8627(.005)	0.5509(.013)	0.5229(.010)	0.4475(.032)	0.8678(.007)	0.5683(.011)	0.5510(.017)	0.4695(.025)
LGTRL-DE	✓	✓	✓	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

AUROC and F1-score on the MIMIC-III dataset. INPREM multiplies the output of the linear module and the derived weights of Transformer-based nonlinear module to obtain the final patient representation. The feed-forward network-based linear module in INPREM might be not very effective to learn the deep representations of the temporal data, resulting in the poorer performance (e.g., an AUROC of 0.8557 on the MIMIC-III dataset) than our proposed method. Compared to the four Transformer-based baselines [18,19,21,30] that capture global dependencies, our proposed method achieves better performance in terms of all four metrics, which implies that capturing both local and global dependencies is effective. It is worth noting that ConCare benefits from a multi-head self-attention mechanism to extract interrelationships between features for personalized medical prediction and obtain a more fine-grained representation of the patient. As a result, ConCare achieves an AUROC of 0.8632 on MIMIC-III. Similarly, AdaCare captures the long-term and short-term variation patterns of clinical features based on multi-scale dilated convolution, which obtains AUROC values of 0.8608 and 0.8678 on the MIMIC-III and e-ICU datasets, respectively.

The above results fully demonstrate the strong ability of our method to capture temporal dependencies. On the one hand, this is due to LGTRL-DE can well capture local and global dependencies among clinical events and fully learn patient representations from EMR data. On the other hand, LGTRL-DE effectively combines the dynamic clinical data with static demographic information to achieve more accurate mortality prediction.

5.2. Ablation study

To measure the effectiveness of different components of the model, we conduct ablation studies on two datasets. Specifically, we remove the different modules in turn and evaluate their performances. In addition, we also investigate the impact of demographic data on prediction performance. Our ablation studies are described below.

5.2.1. Effectiveness of LTRLM

To investigate the effectiveness of the local temporal representation learning module, we compare our model with its variant $LGTRL\text{-}DE_{L-}$, which is obtained by removing the LTRLM from the original model and feds the original input P into the next module directly. We can see from Table 3 that when this module is removed, all metrics of LGTRL-DE decrease on both two datasets. Taking the MIMIC-III dataset as an example, the AUROC, AUPRC, min(Se, P+) and F1-score values of $LGTRL\text{-}DE_{L-}$ are decreased by 1.60%, 3.76%, 2.41% and 8.35% respectively. The results show that LTRLM can effectively capture the temporal patterns of patient health status and the local dependencies of clinical events, thus improving the prediction performance of the model.

5.2.2. Effectiveness of GTRLM

To investigate the impact of the global temporal representation learning module, we compare the performance differences of the models with and without GTRLM. As shown in Table 3, the performance of the variant model $LGTRL\text{-}DE_{G-}$ after removing GTRLM is inferior to that of the original one on both two datasets. For example, it is evident from the results on MIMIC-III that the values of AUROC, AUPRC, and min(Se, P+) of $LGTRL\text{-}DE_{G-}$ decreased from 0.8685, 0.5607, and 0.5367 to 0.8605, 0.5452, and 0.5185, respectively. The experimental results show that the Transformer-based GTRLM can help the model understand the relationships between different clinical events from a global perspective and enhance the feature representation ability of the model.

5.2.3. Effectiveness of MRFM

To verify the effectiveness of this module, we construct a comparative model $LGTRL\text{-}DE_{M-}$ obtained by replacing the MRAM in our original model with a simple flatten layer. We can see in Table 3 that the performance of $LGTRL\text{-}DE_{M-}$ decreases on both datasets. For example, $LGTRL\text{-}DE_{M-}$ achieves AUROC values of 0.8627 and 0.8678 for MIMIC-III and e-ICU datasets, respectively, which are both poorer than our proposed method. The experimental results demonstrate the necessity and effectiveness of MRFM.

5.2.4. Effectiveness of demographic information

To explore the impact of demographic information on prediction performance, we set up three variants of LGTRL-DE: $LGTRL\text{-}DE_{D-}$ that uses no demographic information, $LGTRL\text{-}DE_{DM}$ that uses demographic information only in the MRFM, and $LGTRL\text{-}DE_{DL}$ that uses demographic information only in the LTRLM for initializing hidden units. We can see from Table 4 that the performance of models with and without demographic information differs significantly. Taking the results on the MIMIC-III dataset as an example, the model $LGTRL\text{-}DE_{D-}$ without using demographic information only obtains an AUROC of 0.8579, which performs the worst among the three variants. This indicates that demographic information can indeed provide supplements for comprehensively understanding the development of patient health status. In addition, the performance of $LGTRL\text{-}DE_{DL}$ is better than

Table 4
Effect of demographic information on model performance. The number in () indicates the standard deviation.

Methods	MIMIC-III dataset				e-ICU dataset			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
$LGTRL\text{-}DE_{D-}$	0.8579(.007)	0.5372(.018)	0.5162(.019)	0.4021(.021)	0.8653(.006)	0.5661(.018)	0.5448(.017)	0.4228(.041)
$LGTRL\text{-}DE_{DM}$	0.8632(.003)	0.5534(.008)	0.5267(.014)	0.4522(.053)	0.8694(.008)	0.5722(.015)	0.5502(.015)	0.4730(.033)
$LGTRL\text{-}DE_{DL}$	0.8661(.002)	0.5586(.012)	0.5271(.012)	0.4711(.012)	0.8683(.007)	0.5698(.012)	0.5527(.016)	0.4708(.029)
LGTRL-DE	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

that of $LGTRL\text{-}DE_{D-}$, which further proves that initializing the Bi-GRU hidden state units with demographic information can significantly improve the effectiveness of patient representations. Similarly, the results of $LGTRL\text{-}DE_{DM}$ demonstrate the reliability of fusing demographic information in the MRFM. In summary, our method obtains the best performance, indicating that our method can fully consider the impact of demographic information.

In summary, According to results of ablation studies, removing LTRLM, removing demographic information in both LTRLM and MRFM, and removing GTRLM lead to a top-3 decrease in terms of AUROC on the MIMIC-III dataset (1.60%, 1.06%, and 0.80%, respectively), which implies that the LTRLM, demographic information utilization, and GTRLM are the top-3 important components for in-hospital mortality prediction. These results guide how to design model architecture for better in-hospital mortality prediction. First, to learn effective local representation from the time series data, we should pay more attention to the temporal representation learning and local attention. Second, we should design effective operations to use demographic information to guide temporal representation learning and to make full use of the complementarity of demographic to the time series data (i.e., multi-modality data fusion is important). Third, except for effective local representation learning, we should also focus on designing modules to capture effective global temporal representations.

6. Discussion

In the above experiments and ablation studies, we prove the superiority of the proposed model and the effectiveness of each module. In this section, we will discuss the impact of different aggregation and fusion methods in MRFM, the effect of local attention mechanism, the impact of the number of layers in Transformer block and the impact of different lengths of observation windows on the results. Besides, we perform the calibration study on MIMIC-III and e-ICU datasets.

6.1. Exploring different aggregation methods over time steps in MRFM

In our model, we use the Bi-GRU network in the MRFM to aggregate the clinical event vectors of multiple time steps. Some existing methods aggregate the representation vectors over time steps by element-wise operations (average, concatenation) and recurrent neural networks (GRU, Bi-RNN, Bi-LSTM) [41]. To verify the effectiveness of the aggregation method adopted in our paper, we compare the performance of LGTRL-DE when using the different methods mentioned above on two datasets. As shown in Table 5, Bi-GRU obtains the best performance among all the compared methods. It is worth noting that the aggregation approaches applying concatenation and average operations obtain relatively low AUROC values because they fail to effectively capture the temporal patterns of clinical events. Compared to Bi-RNN and Bi-LSTM, Bi-GRU benefits from its more streamlined internal structure and outperforms Bi-RNN and Bi-LSTM in terms of performance, which is consistent with the findings in [42]. In conclusion, our aggregation method can effectively extract contextual semantic associations of clinical events to generate robust unified health representations.

6.2. Selection of fusion methods in MRFM

In the MRFM, we need to fuse the final dynamic temporal features and the static demographic features to obtain the final patient representation. In fact, there are several different vector processing methods when fusing these two features, such as performing element-wise operations (average, multiplication, and summation) or concatenating these two feature vectors. In order to investigate the most suitable fusion method for our model, we conduct experiments on two datasets to evaluate the model performance when using different fusion methods. The results are shown in Table 6, we can observe that the model performance is relatively low when using the average, multiplication, and summation fusion methods. The reason for this result could be that the direct vector computation of two features with different characteristics (dynamic and static) causes a loss of semantic information. On the contrary, using the concatenation operation does not change the original feature information and achieves the best performance on both datasets. Therefore, we choose the concatenation operation as the final method of fusing dynamic and static features for the model.

6.3. Effect of local attention mechanism

In the local temporal representation learning module, we use the local attention mechanism to capture local temporal patterns of clinical events to obtain short-term changes in patient health status. To validate the effectiveness of the method, we compare the performance after removing the local attention mechanism on two datasets. As shown in Table 7, we can see that LGTRL-DE outperforms the model without the local attention mechanism in all metrics. Taking the MIMIC-III dataset as an example, when compared with the original LGTRL-DE model, the values of AUROC, AUPRC, min(Se, P+) and F1-score decrease from 0.8685, 0.5607, 0.5367 and 0.4928 to 0.8634, 0.5515, 0.5246 and 0.4478. This indicates that short-term changes in patient health status are critical for predicting the final clinical outcome. The introduction of the local attention mechanism allows our model to focus on short-term changes in patient health status, thus improving the accuracy of prediction.

6.4. Impact of the number of layers in transformer block

In this study, one of the important hyperparameters is the number of layers stacked in the Transformer block, i.e., L. To explore the optimal value of L, we make $L = 2/3/4/5/6$ and evaluate the performance of the model on two datasets with different values of L, respectively. Taking the AUROC evaluation metric as an example, Fig. 3 shows the performances on the test set when L is taken to different values. From Fig. 3, we can see that the performance of the model is gradually getting better as L increases initially, and the model performance reaches the highest when $L = 4$. However, when $L > 4$, the model performance decreases as the value of L increases. This could be explained by the fact that as L increases, the network depth and the number of parameters increase, and the model might suffer from over-fitting. Therefore, considering the model performance and efficiency, we choose $L = 4$ as the final number of Transformer block.

Table 5
The performance of LGTRL-DE with different aggregation methods in MRFM. The number in () indicates the standard deviation.

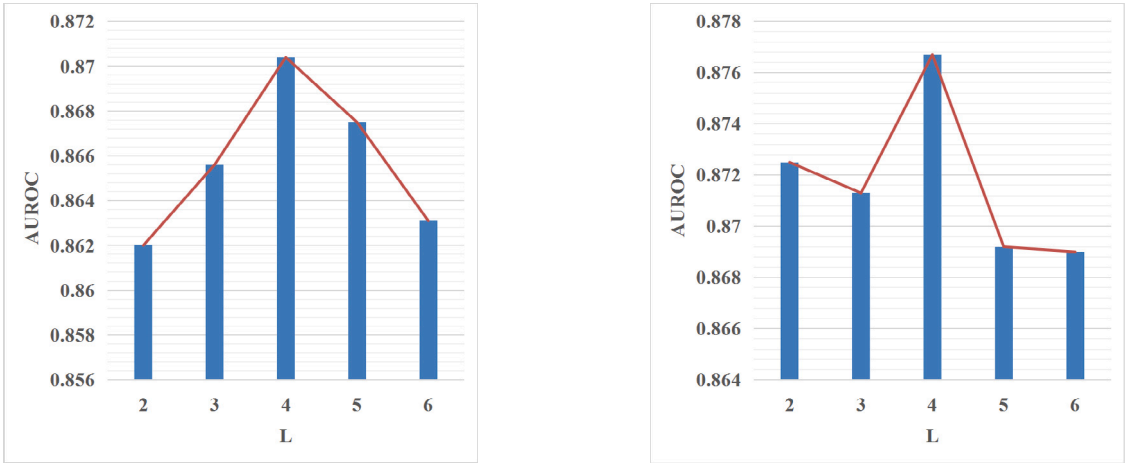
Aggregation methods	MIMIC-III dataset				e-ICU dataset			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
Concatenation	0.8630(.005)	0.5518(.011)	0.5272(.012)	0.4263(.056)	0.8687(.004)	0.5674(.010)	0.5648(.019)	0.4380(.056)
Average	0.8564(.003)	0.5361(.014)	0.5174(.009)	0.4739(.026)	0.8674(.005)	0.5583(.008)	0.5395(.016)	0.4840(.030)
GRU	0.8569(.008)	0.5363(.012)	0.5175(.006)	0.4158(.069)	0.8698(.008)	0.5712(.018)	0.5449(.026)	0.4234(.090)
Bi-RNN	0.8642(.002)	0.5534(.010)	0.5251(.012)	0.4568(.026)	0.8697(.004)	0.5701(.011)	0.5468(.020)	0.4336(.033)
Bi-LSTM	0.8629(.002)	0.5456(.010)	0.5231(.018)	0.4653(.042)	0.8697(.003)	0.5731(.017)	0.5494(.022)	0.4763(.032)
Bi-GRU	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

Table 6
Effect of different fusion methods in MRFM on model performance. The number in () indicates the standard deviation.

Fusion methods	MIMIC-III dataset				e-ICU dataset			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
Multiplication	0.8592(.003)	0.5452(.011)	0.5237(.012)	0.4659(.046)	0.8634(.008)	0.5706(.017)	0.5421(.018)	0.4452(.042)
Average	0.8624(.001)	0.5490(.010)	0.5316(.014)	0.4368(.088)	0.8658(.003)	0.5729(.013)	0.5467(.016)	0.4682(.025)
Summation	0.8637(.006)	0.5523(.019)	0.5265(.017)	0.4169(.043)	0.8677(.005)	0.5733(.015)	0.5504(.019)	0.4576(.038)
Concatenation	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

Table 7
Effect of local attention mechanism. The number in () indicates the standard deviation.

Datasets	Without LAM				With LAM			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
MIMIC-III	0.8634(.003)	.5515(.015)	0.5246(.011)	0.4478(.027)	0.8685(.003)	0.5607(.016)	0.5367(.016)	0.4928(.015)
e-ICU	0.8708(.005)	0.5747(.012)	0.5475(.019)	0.4305(.022)	0.8733(.005)	0.5801(.014)	0.5588(.017)	0.5054(.018)



(a) MIMIC-III (b) e-ICU

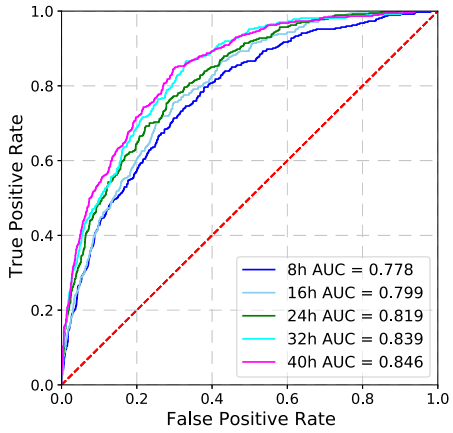
Fig. 3. The impact of the number of layers in Transformer block on the prediction performance on two datasets.

6.5. Impact of different length of observation windows

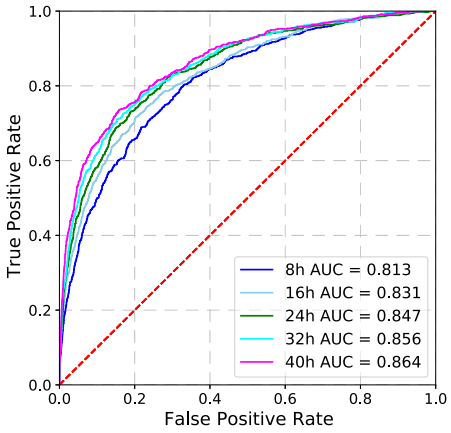
In this study, we evaluate the performance of LGTRL-DE and baseline models using a 48-hour observation window. To explore the effect of different lengths of observation windows, we compare the model performance when the observation window length is 8/16/24/36/40 h, respectively. As shown in Fig. 4, on both datasets, the model trained with a smaller observation window shows slightly lower performance compared with the model trained with a larger observation window. This is because the larger the observation window, the richer the information contained, which leads to a more accurate prediction. In addition, we also compare the performance of $Transformer_e$ when using different lengths of observation windows as described above. From Fig. 4 we can see that compared to $Transformer_e$, our model still maintains good performance when using different lengths of observation windows. It demonstrates that our model can be well adapted even if the observation window length is changed. This benefits from the model's ability to capture local and global dependencies among clinical events.

6.6. Impact of kernel size in CNN

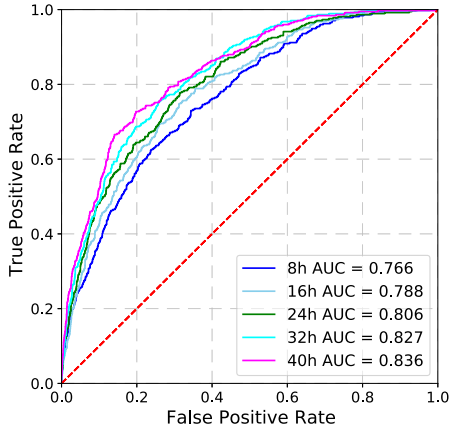
In the LAM module, we apply CNN with a kernel size of 3 and a stride of 1 to capture local information. To explore the effect of different kernel sizes of CNN, we perform experiments using different kernel sizes (1, 3, 5, 7, and 9) in CNN and the results are shown in Table 8. Taking the MIMIC-III dataset as an example, we can find that when we use a kernel size of 3, the proposed method achieves an AUROC of 0.8685 and an AUPRC of 0.5607 which are better than those of using other kernel sizes. The reason can be explained as follows. In general, the change between each clinical event and its two adjacent clinical events is the smallest. Thus, the local correlation between each clinical event and its two adjacent clinical events (i.e., K = 3) is the strongest, and a larger K may weaken the local correlation, making K = 3 more effective.



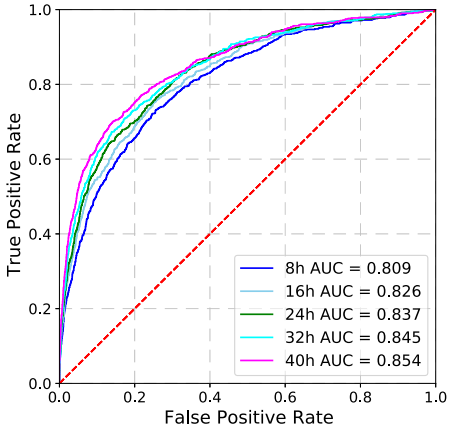
(a) The results of LGTRL-DE on the MIMIC-III dataset



(b) The results of LGTRL-DE on the e-ICU dataset



(c) The results of $Transformer_e$ on the MIMIC-III dataset



(d) The results of $Transformer_e$ on the e-ICU dataset

Fig. 4. Comparison of AUROC curves of LGTRL-DE and $Transformer_e$ for different lengths of observation windows on two datasets.

Table 8
Performance comparison of different kernel sizes in CNN on both datasets.

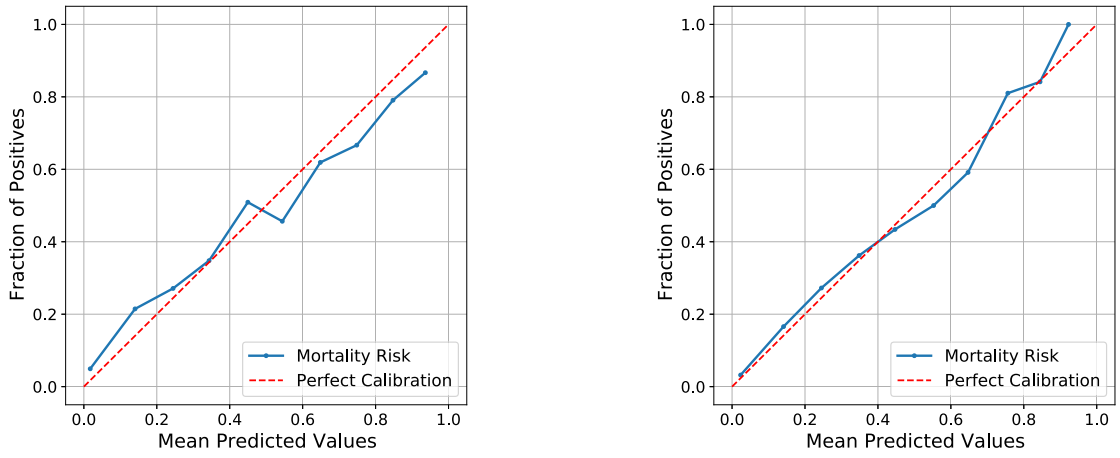
Kernel size	MIMIC-III dataset				e-ICU dataset			
	AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
K = 1	0.8651(.004)	0.5591(.013)	0.5317(.012)	0.4812(.023)	0.8701(.006)	0.5734(.015)	0.5463(.016)	0.4842(.020)
K = 3	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**
K = 5	0.8645(.006)	0.5526(.015)	0.5277(.010)	0.4457(.030)	0.8714(.006)	0.5785(.017)	0.5530(.020)	0.4978(.023)
K = 7	0.8647(.004)	0.5484(.013)	0.5201(.011)	0.4449(.041)	0.8709(.005)	0.5797(.013)	0.5563(.016)	0.4734(.028)
K = 9	0.8643(.008)	0.5425(.016)	0.5235(.011)	0.4847(.028)	0.8710(.005)	0.5715(.008)	0.5482(.018)	0.4704(.025)

6.7. Calibration study

In addition to evaluation metrics such as AUROC and AUPRC, the calibration curve is also a common measure of the model performance, which can help us further judge the reliability of medical models [43, 44]. The probability of the model classifier output indicates the confidence that the sample belongs to a certain category (e.g., positive sample). However, the model's predicted probabilities are biased in practice, so we need to perform probability calibration to make the model's predicted probabilities correspond to the actual probabilities. The closer the calibration curve of a model is to the diagonal, the more accurate the classification prediction of the model is. To this end, we provide calibration curves of the proposed model LGTRL-DE for the in-hospital mortality prediction task on two datasets. As shown in Fig. 5, the calibration curves of the model are essentially close to the diagonal line, which indicates that our model is reliable. On the MIMIC-III dataset, the calibration curve has some fluctuations, and the model is slightly overconfident at lower predicted values and underconfident at higher predicted values.

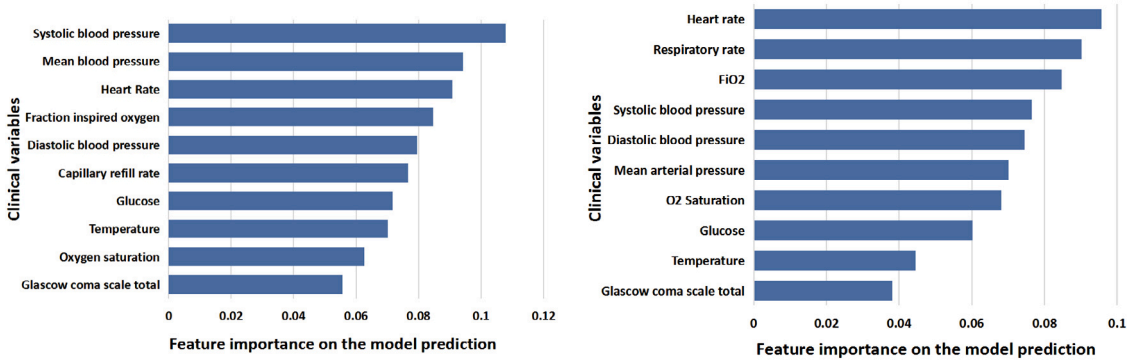
6.8. Analysis of feature importance

To explore the interpretability of the proposed LGTRL-DE, we compute feature importance by calculating the contribution of each feature to the model output for the in-hospital mortality task via using SHapley Additive exPlanation (SHAP) [45] on the MIMIC-III and e-ICU datasets. We first calculate the average of SHAP values for all time steps across all features to obtain the overall impact of each feature on the model output. Then, we take the top-10 most important features, and the results are shown in Fig. 6. The x-axis indicates the mean impact of each feature on the model prediction, and the y-axis indicates the features ranked by importance. Similarly, for the demographic features, we also calculate the SHAP values for each feature on the test set, and the results are shown in Fig. 7.



(a) Calibration curve on the MIMIC-III dataset (b) Calibration curve on the e-ICU dataset

Fig. 5. Calibration curves of LGTRL-DE on two datasets. We set the number of bins to 10 and the curves show the mean mortality predicted values (x-axis) and the actual fraction of positives in each bin (y-axis).



(a) The top-10 clinical features on the MIMIC-III dataset (b) The top-10 clinical features on the e-ICU dataset

Fig. 6. The top-10 clinical time-series features on the MIMIC-III and e-ICU datasets for mortality prediction ranked by the SHAP method.

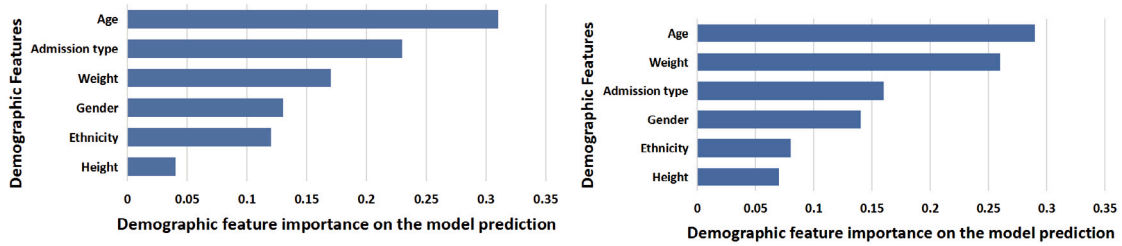
From Fig. 6, it can be seen that 5 out of the top-10 important features (i.e., Systolic blood pressure, Heart rate, Diastolic blood pressure, Glucose, and Glascow coma scale total) are the same between the two datasets. In addition, Systolic blood pressure, Heart rate, and Diastolic blood pressure in the top-5 important features on both datasets are related to blood pressure and heart rate, which point to primary diseases such as hypertension. This indicates that patients with primary diseases such as hypertension have a higher risk of mortality, which is consistent with the study reported in [46]. As seen in Fig. 7, age is the most important demographic feature for mortality, which reflects the clinical fact that older critically ill patients face a higher mortality risk. In addition, weight and admission type also have an impact on mortality as they reflect the patient's base condition (e.g., obesity or not, emergency or not).

6.9. Application of the proposed method to length of stay prediction

In addition to the mortality risk prediction task, we apply the proposed LGTRL-DE method to in-hospital length of stay (LoS) prediction and compared it with other baseline models on the MIMIC-III dataset. To implement LoS prediction, we divide the length of stay into 10 buckets/classes days: 0–1, 1–2, 2–3, 3–4, 4–5, 5–6, 6–7, 7–8, 8–14, and 14+ days, which has been proved to be effective in [4,36]. Besides, we modify our proposed LGTRL-DE and all the compared baselines via only changing the output of the last layer from 1 to 10 and changing the loss function to the sparse categorical cross-entropy loss. We use mean absolute deviation (MAD), mean absolute percentage error (MAPE) and Kappa to evaluate the performance for LoS prediction and the lower the metric is, the better the performance is, except for Kappa. Table 9 shows the results of all methods on the MIMIC-III dataset for LoS prediction. From Table 8, we can see that the proposed LGTRL-DE achieves a MAD of 81.78, a MAPE of 234.66 and a Kappa of 0.4288 outperforming all the baseline methods. The results show that our proposed method is still effective for LoS prediction.

Table 9
Performance of our proposed method and eight baselines on the MIMIC-III dataset for LoS prediction. The number in () indicates the standard deviation over 5-fold cross-validation.

Methods	MIMIC-III dataset		
	MAD	MAPE	Kappa
Bi-GRU	90.09(2.72)	248.62(10.66)	0.4176(0.012)
Retain [26]	90.62(4.38)	274.39(17.47)	0.3808(0.012)
$Transformer_e$ [18]	91.62(3.18)	280.37(24.24)	0.4082(0.014)
SAnd [19]	90.21(6.19)	267.94(32.02)	0.4227(0.016)
STraTs [21]	97.45(5.23)	347.72(55.02)	0.3714(0.015)
INPREM [30]	89.14(8.16)	258.00(23.03)	0.3864(0.036)
ConCare [25]	82.43(4.43)	282.43(23.29)	0.3690(0.012)
AdaCare [17]	90.51(4.27)	245.33(2.64)	0.4154(0.009)
LGTRL-DE	**81.78(3.72)**	**234.66(14.81)**	**0.4288(0.007)**



(a) The feature importance of demographic information on the MIMIC-III dataset

(b) The feature importance of demographic information on the e-ICU dataset

Fig. 7. The feature importance of demographic information on the two datasets for mortality prediction ranked by the SHAP method.

Table 10
Performance of our proposed LGTRL-DE with or without cross-examination on the two datasets. The number in () indicates the standard deviation over 5-fold cross-validation.

Method of testing	Methods	MIMIC-III dataset				e-ICU dataset			
		AUROC	AUPRC	min(Se, P+)	F1-score	AUROC	AUPRC	min(Se, P+)	F1-score
Cross-examination	AdaCare [17] (common features)	0.7928(.006)	0.4236(.004)	0.4479(.012)	0.3261(.027)	0.8227(.005)	0.4229(.006)	0.4467(.011)	0.3265(.024)
	LGTRL-DE (common features)	0.8269(.002)	0.4517(.008)	0.4807(.010)	0.3452(.019)	0.8471(.004)	0.4512(.008)	0.4843(.010)	0.3428(.021)
w/o Cross-examination	LGTRL-DE (common features)	0.8385(.003)	0.4602(.016)	0.4875(.014)	0.3642(.039)	0.8503(.006)	0.5132(.013)	0.4921(.013)	0.4078(.023)
	LGTRL-DE (original features)	**0.8685(.003)**	**0.5607(.016)**	**0.5367(.016)**	**0.4928(.015)**	**0.8733(.005)**	**0.5801(.014)**	**0.5588(.017)**	**0.5054(.018)**

6.10. The generalizability of the proposed method

To cross-examine the generalizability of the proposed method, we select 12 common clinical features and 6 demographic features on the MIMIC-III and e-ICU datasets to retrain models on the source dataset and then test the models on the target dataset. Specifically, we compare the proposed LGTRL-DE model and the best baseline model (AdaCare) by training both models on the MIMIC-III dataset in a 5-fold cross-validation manner and testing them on the whole e-ICU dataset and vice versa. In addition, we compare the performance of the proposed method using common features without cross-examination. Experimental results are shown in Table 10. We can see that in cross-examination scenarios, our proposed LGTRL-DE still outperforms the best baseline for in-hospital mortality prediction and achieves good performance close to that of our proposed LGTRL-DE using common features without cross-examination, showing the proposed method has good generalizability. Besides, in direct examination scenarios, because the number of common features is smaller than the number of original features, our proposed LGTRL-DE with common features is poorer than our proposed method with original features, which demonstrates the used feature selection method is necessary and effective.

6.11. Limitation and future work

Deep learning methods based on EMRs data show great prospect in in-hospital mortality prediction task. Although LGTRL-DE achieves better performance, it can only handle temporal data, and does not take into account other rich textual data (e.g., physician notes) and dynamically recorded signal data (e.g., ECG) in the EHRs. Future work will focus on better incorporation of data from other modalities to achieve more accurate mortality prediction. In addition, LGTRL-DE implicitly assumes that the time interval between different clinical events is the same and does not consider the influence of irregular time, which may lose part of the patient's status information. In future work, we will introduce the method of processing irregular time-series data to solve this problem.

7. Conclusion

In this work, we propose LGTRL-DE, a novel deep representation learning framework for in-hospital mortality prediction. Specifically, we generate the final patient representation for prediction by modeling dynamic clinical time series with static demographic information. The proposed approach extracts temporal information of clinical time series from both local and global perspectives and effectively combines temporal information and static data. We evaluate LGTRL-DE on two public EMR datasets: MIMIC-III and e-ICU. Experimental results show that our model effectively improves the prediction performance and outperforms several state-of-the-art methods.

CRediT authorship contribution statement

Mengjie Zou: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Ying An:** Conceptualization, Investigation, Methodology, Formal analysis, Data curation, Writing – review & editing, Funding acquisition. **Hulin Kuang:** Conceptualization, Investigation, Methodology, Formal analysis, Writing – review & editing. **Jianxin Wang:** Conceptualization, Validation, Writing – review & editing, Resources, Funding acquisition, Supervision.

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.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2021YFF1201200), the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (No. U1909208), and the Science and Technology Major Project of Changsha (No. kh2202004). This work was also carried out in part using computing resources at the High Performance Computing Center of Central South University.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104408.

References

[1] Y. Yin, C.-A. Chou, A novel switching state-space model for post-ICU mortality prediction and survival analysis, IEEE J. Biomed. Health Inf. 25 (9) (2021) 3587–3595.

[2] K. Lin, Y. Hu, G. Kong, Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model, Int. J. Med. Inform. 125 (2019) 55–61.

[3] C. Steinmeyer, L. Wiese, Sampling methods and feature selection for mortality prediction with neural networks, J. Biomed. Inform. 111 (2020) 103580.

[4] E. Rocheteau, P. Liò, S. Hyland, Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 58–68.

[5] Y. Feng, Z. Xu, L. Gan, N. Chen, B. Yu, T. Chen, F. Wang, DCMN: Double core memory network for patient outcome prediction with multimodal data, in: 2019 IEEE International Conference on Data Mining, ICDM, IEEE, 2019, pp. 200–209.

[6] L. Liu, H. Li, Z. Hu, H. Shi, Z. Wang, J. Tang, M. Zhang, Learning hierarchical representations of electronic health records for clinical outcome prediction, in: AMIA Annual Symposium Proceedings, Vol. 2019, American Medical Informatics Association, 2019, p. 597.

[7] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, L. Jorm, Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk, Sci. Rep. 10 (1) (2020) 1–10.

[8] Y. An, N. Huang, X. Chen, F. Wu, J. Wang, High-risk prediction of cardiovascular diseases via attention-based deep neural networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (3) (2019) 1093–1105.

[9] L. Cui, S. Biswal, L.M. Glass, G. Lever, J. Sun, C. Xiao, CONAN: complementary pattern augmentation for rare disease detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 01, 2020, pp. 614–621.

[10] Y. Zhang, ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI-2019, Macao, China, 2019, pp. 4369–4375.

[11] M. Critical Data, Secondary Analysis of Electronic Health Records, Springer Nature, 2016.

[12] K. Yu, Z. Yang, C. Wu, Y. Huang, X. Xie, In-hospital resource utilization prediction from electronic medical records with deep learning, Knowl.-Based Syst. 223 (2021) 107052.

[13] E. Jun, A.W. Mulyadi, J. Choi, H.-I. Suk, Uncertainty-gated stochastic sequential model for EHR mortality prediction, IEEE Trans. Neural Netw. Learn. Syst. 32 (9) (2020) 4052–4062.

[14] S. Baker, W. Xiang, I. Atkinson, Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach, Sci. Rep. 10 (1) (2020) 1–12.

[15] I. Gandin, A. Scagnetto, S. Romani, G. Barbati, Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to Intensive care unit, J. Biomed. Inform. 121 (2021) 103876.

[16] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, PMLR, 2013, pp. 1310–1318.

[17] L. Ma, J. Gao, Y. Wang, C. Zhang, J. Wang, W. Ruan, W. Tang, X. Gao, X. Ma, Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 01, 2020, pp. 825–832.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[19] 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.

[20] Y. Li, S. Rao, J.R.A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: transformer for electronic health records, Sci. Rep. 10 (1) (2020) 1–12.

[21] S. Tipirneni, C.K. Reddy, Self-supervised transformer for multivariate clinical time-series with missing values, 2021, arXiv preprint arXiv:2107.14293.

[22] C. Delon, K.F. Brown, N.W. Payne, Y. Kotrotsios, S. Vernon, J. Shelton, Differences in cancer incidence by broad ethnic group in England, 2013–2017, Br. J. Cancer (2022) 1–9.

[23] G. Levy, N. Schupf, M.-X. Tang, L.J. Cote, E.D. Louis, H. Mejia, Y. Stern, K. Marder, Combined effect of age and severity on the risk of dementia in Parkinson's disease, Ann. Neurol. Official J. Am. Neurol. Assoc. Child Neurol. Soc. 51 (6) (2002) 722–729.

[24] J. Gao, C. Xiao, Y. Wang, W. Tang, L.M. Glass, J. Sun, Stagenet: Stage-aware neural networks for health risk prediction, in: Proceedings of the Web Conference 2020, 2020, pp. 530–540.

[25] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, J. Gao, Concare: Personalized clinical feature embedding via capturing the healthcare context, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 01, 2020, pp. 833–840.

[26] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. 29 (2016).

[27] G. Harerimana, J.W. Kim, B. Jang, A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data, J. Biomed. Inform. 118 (2021) 103778.

[28] S. Wang, J. Liu, ClinicNet: Clinical practice oriented medical representation learning for electronic medical records, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2020, pp. 2097–2104.

[29] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1903–1911.

[30] X. Zhang, B. Qian, S. Cao, Y. Li, H. Chen, Y. Zheng, I. Davidson, INPREM: An interpretable and trustworthy predictive model for healthcare, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 450–460.

[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[32] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.

[33] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[35] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[36] 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.

[37] T. Na Pattalung, T. Ingviya, S. Chaichulee, Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients, J. Pers. Med. 11 (9) (2021) 934.

[38] T.J. Pollard, A.E. Johnson, J.D. Raffa, L.A. Celi, R.G. Mark, O. Badawi, The eICU Collaborative Research Database, a freely available multi-center database for critical care research, Sci. Data 5 (1) (2018) 1–13.

[39] S. Sheikhalishahi, V. Balaraman, V. Osmani, Benchmarking machine learning models on multi-centre eICU critical care dataset, PLoS One 15 (7) (2020) e0235424.

[40] I. Silva, G. Moody, D.J. Scott, L.A. Celi, R.G. Mark, Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012, in: 2012 Computing in Cardiology, IEEE, 2012, pp. 245–248.

[41] R. Weng, H. Wei, S. Huang, H. Yu, L. Bing, W. Luo, J. Chen, Gret: Global representation enhanced transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, 2020, pp. 9258–9265.

[42] B. Bardak, M. Tan, Improving clinical outcome predictions using convolution over medical entities with multimodal learning, Artif. Intell. Med. 117 (2021) 102112.

[43] A.C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. Devereaux, T. McGinn, G. Guyatt, Discrimination and calibration of clinical prediction models: users' guides to the medical literature, JAMA 318 (14) (2017) 1377–1384.

[44] X. Jiang, M. Osl, J. Kim, L. Ohno-Machado, Calibrating predictive model estimates to support personalized medicine, J. Am. Med. Inform. Assoc. 19 (2) (2012) 263–274.

[45] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[46] N.R. Poulter, D. Prabhakaran, M. Caulfield, Hypertension, Lancet 386 (9995) (2015) 801–812.