DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series

Qingxiong Tan,[1] Mang Ye,[1] Baoyao Yang,[1] Si-Qi Liu,[1] Andy Jinhua Ma,[2]
Terry Cheuk-Fung Yip,[3] Grace Lai-Hung Wong,[3] Pong C. Yuen[1]*
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
[3]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong
{csqxtan, mangye, byyang, siqiliu, pcyuen}@comp.hkbu.edu.hk,
{andyjinhuama, terryfungyip}@gmail.com, wonglaihung@cuhk.edu.hk

Abstract

Due to the discrepancy of diseases and symptoms, patients usually visit hospitals irregularly and different physiological variables are examined at each visit, producing large amounts of irregular multivariate time series (IMTS) data with missing values and varying intervals. Existing methods process IMTS into regular data so that standard machine learning models can be employed. However, time intervals are usually determined by the status of patients, while missing values are caused by changes in symptoms. Therefore, we propose a novel end-to-end Dual-Attention Time-Aware Gated Recurrent Unit (DATA-GRU) for IMTS to predict the mortality risk of patients. In particular, DATA-GRU is able to: 1) preserve the informative varying intervals by introducing a time-aware structure to directly adjust the influence of the previous status in coordination with the elapsed time, and 2) tackle missing values by proposing a novel dual-attention structure to jointly consider data-quality and medical-knowledge. A novel unreliability-aware attention mechanism is designed to handle the diversity in the reliability of different data, while a new symptom-aware attention mechanism is proposed to extract medical reasons from original clinical records. Extensive experimental results on two real-world datasets demonstrate that DATA-GRU can significantly outperform state-of-the-art methods and provide meaningful clinical interpretation.

Introduction

The widely-used electronic health records (EHR) produces a large quantity of health data, providing valuable opportunities to develop advanced machine learning methods to improve healthcare service (Shickel et al. 2017; Liu et al. 2018a). One important task is to predict the mortality risk of patients based on their historical EHR data. Accurate prediction results help doctors evaluate early treatment effects and design effective treatment plans (Liu et al. 2018b).

This task is challenging since EHR data consists of irregular multivariate time series (IMTS), as illustrated in Fig.1. At different stages of diseases, patients visit hospitals under varying intervals due to the dynamics of health status. Moreover, different physiological variables are examined at different visits because of the changes in symptoms, e.g., when a certain symptom disappears, corresponding variables are no longer examined, resulting in missing values. The ***varying time intervals*** between successive visits and ***missing values*** raise two key challenges in using IMTS to achieve accurate mortality risk prediction results. In addition, these properties provide valuable information in improving prediction performance since they usually reflect the health status and symptoms of patients.

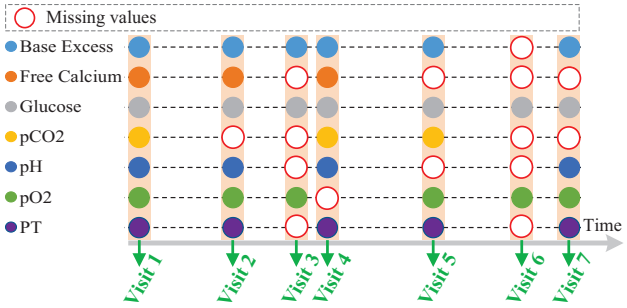


Figure 1: Illustration of irregular multivariate time series (IMTS). The time intervals between successive visits vary significantly from each other. Meanwhile, certain physiological variables are not examined at some visits, causing missing values.

Machine learning methods have been successfully applied in many areas, e.g., speech recognition (Afouras et al. 2018), computer version (Ye et al. 2019a; 2019b; 2019c; Fu et al. 2019), natural language processing (Camburu et al. 2018), energy prediction (Yuan et al. 2017; 2015; Kang et al. 2017). The state-of-the-art sequence modeling methods are recurrent neural networks (RNNs) (Chung et al. 2014; Pang et al. 2019; Gao et al. 2019). The major limitation of standard RNN is that it is designed for data with constant intervals (Baytas et al. 2017), which cannot handle the irregular time-series data. Most clinical prediction methods convert IMTS into equally spaced by discretizing the time axis into non-overlapping intervals with a hand-designed interval (Tan et al. 2019; Xu et al. 2018; Tan et al. 2018; Che et al. 2017; Lipton, Kale, and Wetzel 2016). Then missing values are filled via imputation methods. However, when

the manually selected interval is long, it may cause the loss of temporal information; Conversely, it may increase the missing data rate when the interval is short. Thus, a learning-based method is introduced to obtain an optimal interval in InterpNet (Shukla and Marlin 2019). However, when InterpNet finally specifies the interval, it still unavoidably introduces additional noise or causes information loss, because different patients could have very different numbers of visits. Moreover, these methods usually assume that there is an expected fixed time interval. This assumption may not valid in practice due to the dynamics of diseases.

A better way to handle IMTS data is to directly model the unequally spaced data. Time-aware LSTM (T-LSTM) incorporates irregular time intervals to adjust the hidden status in the memory cell (Baytas et al. 2017). However, T-LSTM is designed for ICD-9 codes, which cannot address the missing data problem in real-valued variables. The recently proposed GRU-D tries to handle both problems (Che et al. 2018). GRU-D introduces observed records and corresponding timestamps into standard GRU to impute missing values as the decay of previous input values toward the overall mean over time. However, GRU-D only combines the empirical mean value and the previous observation to impute missing values. This strategy cannot capture the global structure information of sequence data. Furthermore, GRU-D ignores the diversity in the reliability of different data points, especially the relatively larger unreliability of imputed records compared with actual records. As a result, it assigns equal weights to actually observed data and imputed data, which seriously damages its performance.

To address the aforementioned challenges, this paper presents a novel end-to-end Dual-Attention Time-Aware Gated Recurrent Unit (DATA-GRU) for IMTS to improve the mortality risk prediction performance. To preserve informative varying intervals, which reflect dynamics in the conditions of patients, we introduce a time-aware structure to handle irregular time intervals. This strategy avoids processing IMTS into equally spaced, thus protecting temporal information in dense records and avoiding introducing extra noise to sparse records. Since missing values cause data misalignment, they need to be imputed so as to compose tensor (Comon 2014). However, the imputation process would impair medical information contained in missing values. Therefore, we propose a novel dual-attention structure with two new attention mechanisms to simultaneously focus on the ***data-quality view*** and the ***medical-knowledge view***. For the ***data-quality view***, a novel unreliability-aware attention mechanism is proposed to estimate diversity in the unreliability of different data and accordingly assign them learnable attention weights to ensure high-quality data play more important roles. Our main ideas are that imputed data normally are less reliable than actual records and different imputed data could have different degrees of unreliability, e.g., data inferred from sparse observations are less reliable than from dense observations. For the ***medical-knowledge view***, a novel symptom-aware attention mechanism is proposed to directly extract medical information from original clinical records. Different from other domains, missing values in EHR data possess important medical considerations. It should be noted that DATA-GRU is designed in an end-to-end architecture to ensure the parameters of different parts are trained jointly to achieve global optimal.

The main contributions of this paper are listed as follows:

- We propose a new end-to-end DATA-GRU network with two novel structures to handle the two key challenges in medical IMTS data analysis.
- We introduce a time-aware structure into deep learning architecture to directly incorporate irregular time intervals to adjust the influence of the previous status. This strategy preserves the contained useful information on dynamic changes in the health status of patients.
- We design a new dual-attention structure to handle missing values from both data-quality and medical-knowledge views. Novel unreliability-aware attention is proposed to assign learnable weights to different data in coordination with their reliabilities, while new symptom-aware attention is designed to learn medical information from the sampling characteristics of original EHR data.
- We empirically show that DATA-GRU outperforms state-of-the-art methods on two real-world datasets. The case study indicates that the learned attention weights can provide meaningful clinical interpretation.

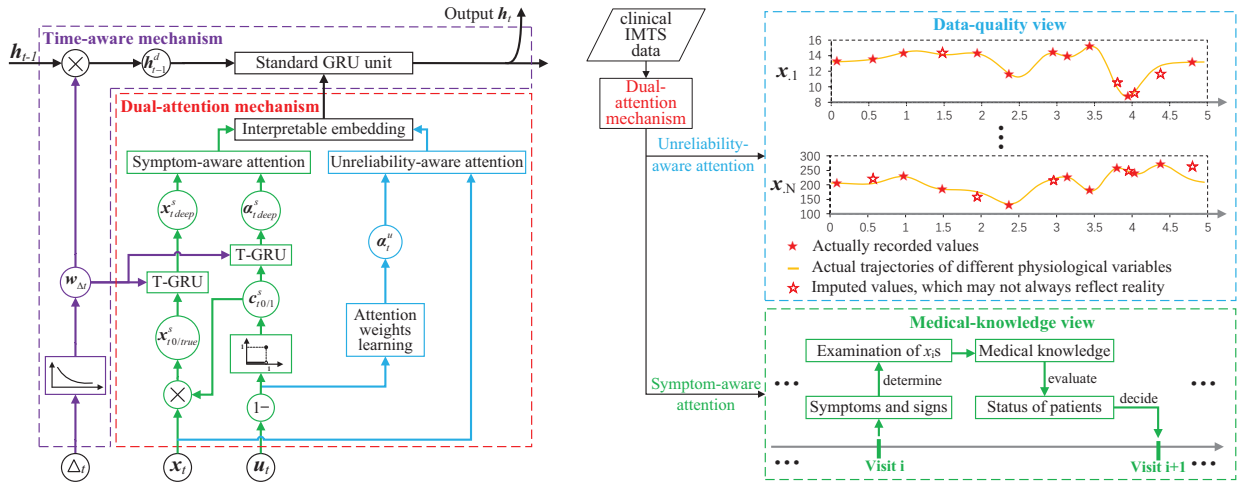
Related Work

Attention methods have been successfully applied in many tasks, e.g., machine translation (Shankar and Sarawagi 2019) and computer vision (Fu et al. 2019). However, data in their domains usually have regular time intervals or have no time attribute, which is unsuitable for the EHR data.

Several works have investigated the attention mechanism for EHR data. To increase the interpretability of networks, RETAIN introduces an attention network to detect influential visits and key variables (Choi et al. 2016). A graph-based attention model is proposed to learn robust representations of EHR data (Choi et al. 2017). Similarly, (Ma et al. 2018) introduce a knowledge-based attention mechanism to embed nodes in the knowledge graph. Three attention mechanisms are introduced to measure relationships between current status and past state in RNN (Ma et al. 2017) and monitor health conditions (Suo et al. 2017). The attention mechanism is also used to handle the low measuring quality problem (Heo et al. 2018). (Song et al. 2018) use masked self-attention to dispense the recurrence in the network. The attention mechanisms used in these methods can promote the performance and interpretability of models at some extents. However, these attention-based methods are usually designed for regular time-series data (or generated regular data), thus cannot be applied to the irregularly sampled EHR data, which is a key problem for health data.

Proposed Method

In this section, we present the proposed Dual-Attention Time-Aware Gated Recurrent Unit (DATA-GRU). We firstly introduce the notations used in this paper.



(a) Architecture of the proposed DATA-GRU (b) Illustration of the dual attention mechanism

Figure 2: Architecture of DATA-GRU and illustration of dual-attention mechanism. (a) DATA-GRU takes input records $\boldsymbol{x}_t$, time intervals Δ_t and unreliability scores $\boldsymbol{u}_t$ as inputs. Parts in purple, green, and cyan denote time-aware mechanism, symptom-aware attention, and unreliability-aware attention, respectively. DATA-GRU handles irregular intervals by converting them into weights via a decay function to adjust the influence of previous status on current status. (b) DATA-GRU tackles missing values by designing a novel dual-attention structure to jointly consider data-quality and medical-knowledge.

Notations

Let $D = \{(\boldsymbol{X}^k, y^k)|k = 1, 2, \cdots, K\}$ denote a dataset with K samples. Each data sample contains a multivariate time series (MTS) and a target value. We represent MTS of the kth sample as $\boldsymbol{X}^k = (\boldsymbol{x}_1^k, \cdots, \boldsymbol{x}_t^k, \cdots, \boldsymbol{x}_{T_k}{}^k) \in R^{T_k \times N}$, where N is the number of all input features; T_k is the number of visits in the kth sample; $\boldsymbol{x}_t^k (1 \leq t \leq T_k)$ denotes records of input features at time t; $\boldsymbol{x}_{t.n}{}^k$ is the value of the nth feature of $\boldsymbol{x}_t^k$. The two key challenges for irregular MTS (IMTS) are: a) time irregularities, which refers to varying intervals between successive visits, i.e., the time intervals dataset $\{t_t - t_{t-1}|t = 1, 2, \cdots, T_k\}$ has multiple values; b) missing values (i.e., several physiological variables are not examined at some visits because of the changes in symptoms), due to which the length of each time-series feature is often shorter than T_k. Thus, we represent $\boldsymbol{X}^k$ as a time series $\boldsymbol{t}_{all}^k$ and a tuple $\boldsymbol{S}^k = (\boldsymbol{t}^k, \boldsymbol{x}^k)$, where $\boldsymbol{t}_{all}^k = [\boldsymbol{t}_1^k, \boldsymbol{t}_2^k, \cdots, \boldsymbol{t}_{T_k}{}^k]$ is the list of timestamps of all visits; $\boldsymbol{t}^k = [\boldsymbol{t}_{.1}^k, \cdots, \boldsymbol{t}_{.n}^k, \cdots, \boldsymbol{t}_{.N}^k]$ is the list of timestamps for each input feature ($\boldsymbol{t}_{.n}^k (1 \leq k \leq K)$ is a subset of $\boldsymbol{t}_{all}^k$); and $\boldsymbol{x}^k = [\boldsymbol{x}_{.1}^k, \cdots, \boldsymbol{x}_{.n}^k, \cdots, \boldsymbol{x}_{.N}^k]$ is the corresponding list of recorded input features.

Dual-Attention Time-Aware GRU (DATA-GRU)

The architecture of the proposed DATA-GRU is shown in Fig.2(a). Compared with standard GRU, DATA-GRU has two novel components, i.e., time-aware mechanism and dual-attention mechanism, which simultaneously handles varying intervals and missing values. The time-aware mechanism is introduced to handle irregular intervals. Although standard GRU (and other variations of RNN) has recursive formulation and can handle variable-length sequences, but it can only handle the time-series data with equal intervals between successive elements. Thus, standard GRU has no structure to handle irregular intervals with missing values. However, as mentioned above, processing IMTS data of different patients into equally spaced would seriously damage data quality, making it nearly impossible to achieve accurate prediction results. Therefore, we introduce the time-aware structure to directly handle varying time intervals.

The dual-attention structure is designed to handle missing values from both data-quality and medical-knowledge views, as illustrated in Fig.2(b). It is noteworthy that imputed values may not always reflect reality and are less reliable than actual records. To achieve accurate and reliable predictions, it is suggested to assign smaller weights to less reliable data and weaken its role by developing an unreliability-aware attention mechanism. In addition, unlike other domains, missing values in EHR data may contain important medical considerations, e.g., whether certain physiological variables are examined or not may indicate the emergence of specific symptoms and signs. Mining such medical information will further promote final results. Therefore, a novel symptom-aware attention is proposed to learn medical information from the medical-knowledge view. The details of DATA-GRU are presented in the following sections.

Time-aware mechanism. As mentioned above, we only impute missing values rather than generating equally spaced data to avoid noise generation. Most methods utilize simple imputation methods, e.g., forward filling with past values and mean value imputation, or their combinations to handle missing values. However, these methods are incapable of capturing the global structure of time-series data. Since Gaussian process (GP) can incorporate global structure information from all available records to conduct imputation, we use it to fill missing values in IMTS. For the kth sample, time series pairs of many variables in the tu-

ple $\boldsymbol{S}^k = (\boldsymbol{t}^k, \boldsymbol{x}^k)$, where $\boldsymbol{t}^k = [\boldsymbol{t}^k_{.1}, \cdots, \boldsymbol{t}^k_{.n}, \cdots, \boldsymbol{t}^k_{.N}]$ and $\boldsymbol{x}^k = [\boldsymbol{x}^k_{.1}, \cdots, \boldsymbol{x}^k_{.n}, \cdots, \boldsymbol{x}^k_{.N}]$, are shorter than the list of all timestamps $\boldsymbol{t}^k_{all} = [\boldsymbol{t}^k_1, \boldsymbol{t}^k_2, \cdots, \boldsymbol{t}_{T_k}{}^k]$ due to missing values. For each variable, we represent actually observed records as $\boldsymbol{X}^k_n$ and corresponding timestamps as $\boldsymbol{T}^k_n$. The estimation of the missing value at time $t_*(t_* \in \boldsymbol{t}^k_{all}\ \&\ t_* \notin \boldsymbol{T}^k_n)$ is made via the computation of the conditional distribution, which is a Gaussian distribution with a mean function $E[x_*]$ and a covariance function $Con[x_*]$:

$$p(x_*|t_*, \boldsymbol{X}^k_n, \boldsymbol{T}^k_n) \sim N(u_*, \delta^2_*) \tag{1}$$

$$E[x_*] = k(t_*)^T (K(\boldsymbol{T}^k_n, \boldsymbol{T}^k_n) + \delta^2 \boldsymbol{I}_N)^{-1} \boldsymbol{X}^k_n \tag{2}$$

$$Con[x_*] = k(t_*, t_*) - k(t_*)^T (K(\boldsymbol{T}^k_n, \boldsymbol{T}^k_n) + \delta^2 \boldsymbol{I}_N)^{-1} k(t_*) \tag{3}$$

where $K(\boldsymbol{T}^k_n, \boldsymbol{T}^k_n)$ is the covariance matrix between observed records; $k(t_*)$ is the covariance matrix between the estimated value and observed records; $\boldsymbol{I}_N$ is a unit matrix.

It should be noted that imputed values may not always be reliable: 1) compared with actual records, imputed values are relatively less reliable since they are inferred from actual records; 2) different imputed data typically have different degrees of reliability. When imputing the missing value at a timestamp with intensive observations nearby, the estimated value is reliable due to the increase of posterior knowledge. Conversely, data inferred from sparse observations are less reliable. GP naturally provides covariance functions to quantitatively describe unreliability of estimated data. Since actual records are relatively absolutely reliable compared with imputed values, we set their unreliability scores to zero. Thus, we get unreliability scores of different data:

$$u[x_*] = \begin{cases} 0, \text{for actually observed records} \\ Con[x_*] > 0, \text{for imputed values} \end{cases} \tag{4}$$

For each sample, we represent its IMTS as three data streams: an augmented input record, i.e., $\boldsymbol{X}^k_r = [\boldsymbol{x}^k_{.1}, \cdots, \boldsymbol{x}^k_{.n}, \cdots, \boldsymbol{x}^k_{.N}] = [\boldsymbol{x}^k_1, \cdots, \boldsymbol{x}^k_t, \cdots, \boldsymbol{x}^k_{T^k}]^T$, which is a unequally spaced MTS without missing values; an unreliability scores matrix $\boldsymbol{U}^k_r = [\boldsymbol{u}^k_{.1}, \cdots, \boldsymbol{u}^k_{.n}, \cdots, \boldsymbol{u}^k_{.N}] = [\boldsymbol{u}^k_1, \cdots, \boldsymbol{u}^k_t, \cdots, \boldsymbol{u}^k_{T^k}]^T$, which has the same shape with $\boldsymbol{X}^k_r$ and quantitatively describes the unreliability degree of each element in $\boldsymbol{X}^k_r$; and a list of time intervals $\Delta^k_{all} = [\Delta^k_1, \cdots, \Delta^k_t, \cdots, \Delta^k_{T_k}]$, where $\Delta_t = t_t - t_{t-1}$.

DATA-GRU takes input records $\boldsymbol{x}_t$, time intervals Δ_t and unreliability scores $\boldsymbol{u}_t$ as inputs, as shown in Fig.2. The time intervals are directly incorporated into DATA-GRU to adjust the hidden status in the previous memory cell. To ensure the influence of previous status fades with the increase of the time interval, we suggest to utilize a decay function to transform it into weight. We tested several decay functions, e.g., $w_{\Delta t=1/log(e+\Delta t)}$, $w_{\Delta t} = e^{-\Delta t}$ and $w_{\Delta t=1/\Delta t}$, and found that $w_{\Delta t=1/log(e+\Delta t)}$ is slightly better. Therefore, we use it to transform time intervals into proper weights to adjust hidden state. The mathematical formulations for $w_{\Delta t}$ and h^d_{t-1} are as follows:

$$w_{\Delta t=1/log(e+\Delta t)} \tag{5}$$

$$h^d_{t-1} = h_{t-1} \odot w_{\Delta t} \tag{6}$$

For convenience, we name the variant of GRU equipped with the time-aware mechanism as T-GRU. Compared with standard GRU, T-GRU can directly analyze unequally spaced univariate or multivariate time series without the necessity of processing it into equally spaced data and thus can preserve the informative varying intervals.

Dual-attention mechanism. Imputed records in augmented data may not always reflect reality and the imputation process could damage medical considerations behind sampling characteristics of original EHR data, both affecting risk prediction. To this end, a novel dual-attention structure is further integrated into T-GRU to handle missing values by jointly considering data quality and medical knowledge.

Unreliability-aware attention is proposed from the data-quality view. Since the degrees of unreliability diverse between actual records and imputed records, and also vary among different imputed records, we propose an unreliability-aware attention mechanism to adjust weights assigned to different data to ensure high-quality data play important roles to promote prediction performance while the influence of low-quality data is limited. For convenience, unreliability score is converted into reliability score via $c_t = 1 - u_t$. Since c_t is only able to identify the quality of different elements within each time series but is unable to identify important variables, we learn unreliability-aware weights from c_t using $\alpha^u_t = sigmoid(W^u c_t + b^u)$ and utilize the learned weights to adjust scores contributed by different elements in time series of different variables. The expressions are given below:

$$c_t = 1 - u_t \tag{7}$$

$$\alpha^u_t = sigmoid(W^u c_t + b^u) \tag{8}$$

$$x^u_t = x_t \odot \alpha^u_t \tag{9}$$

The sampling characteristic of original EHR data possesses important medical considerations. We avoid damaging informative varying intervals in IMTS by introducing the time-aware structure, which is a big step forward compared with the typical methods of processing IMTS into equally spaced. However, the imputation process may still damage some medical information, i.e., missing values which are typically caused by changes in the symptoms of patients. To this end, from the medical-knowledge view, we propose novel symptom-aware attention to further supplement unreliability-aware attention. To exclude the impact of imputed records, we filter out all the imputed values with an actual records pass filter (ARPF), which only allows actually observed records to pass through, namely $c^s_{t\,0/1} = F_{ARPF}(c_t) = \lfloor c_t - 0.5 \rfloor$ for reliability scores and $x^s_{t\,0/true} = x_t \odot c^s_{t\,0/1}$ for input records, such that sampling characteristics of original EHR data are preserved. The filtered data has severe irregularities and the contained medical information is difficult to extract by using standard machine learning methods, whose architectures are designed for regular data. Therefore, we utilize the aforementioned T-GRU to handle the time irregularity problem to extract deep symptom-aware input values $x^s_{t\,deep} = TGRU(x^s_{t\,0/true}, w_{\Delta t})$ and deep symptom-aware attention weights $\alpha^s_{t\,deep} = TGRU(\alpha^s_{t\,0/1}, w_{\Delta t})$. Then, $\alpha^s_{t\,deep}$ are

Table 1: The AUC scores ($mean \pm std$) of different levels of mortality risk predictions for MIMIC-III. **Red** represents the best performance while **Blue** and **Green** indicate the second and third best performance, respectively.

Models	In-hospital	1 day	5 days	10 days	15 days	20 days
LR (Hosmer et al. 2013)	0.815 ± 0.007	0.821 ± 0.009	0.810 ± 0.008	0.798 ± 0.008	0.791 ± 0.007	0.777 ± 0.007
RF (Breiman 2001)	0.849 ± 0.007	0.856 ± 0.009	0.842 ± 0.008	0.834 ± 0.007	0.823 ± 0.007	0.820 ± 0.007
IndRNN (Li et al. 2018)	0.888 ± 0.006	0.894 ± 0.007	0.888 ± 0.007	0.875 ± 0.006	0.867 ± 0.006	0.858 ± 0.006
GRU-raw (Chung et al. 2014)	0.885 ± 0.005	0.888 ± 0.007	0.874 ± 0.007	0.867 ± 0.006	0.857 ± 0.006	0.852 ± 0.006
GRU (Chung et al. 2014)	0.883 ± 0.006	0.892 ± 0.007	0.881 ± 0.006	0.875 ± 0.006	0.866 ± 0.006	0.859 ± 0.006
T-LSTM (Baytas et al. 2017)	0.863 ± 0.006	0.896 ± 0.007	0.872 ± 0.007	0.856 ± 0.007	0.851 ± 0.006	0.840 ± 0.006
T-LSTM-ND (Baytas et al. 2017)	0.885 ± 0.005	0.899 ± 0.007	0.890 ± 0.006	0.874 ± 0.006	0.863 ± 0.006	0.855 ± 0.006
GRU-D (Che et al. 2018)	0.900 ± 0.005	0.923 ± 0.006	0.895 ± 0.006	0.875 ± 0.006	0.876 ± 0.006	0.866 ± 0.006
InterpNet (Shukla and Marlin 2019)	**0.903 ± 0.005**	**0.925 ± 0.005**	**0.901 ± 0.005**	**0.887 ± 0.005**	**0.879 ± 0.005**	**0.872 ± 0.005**
T-GRU-raw	0.889 ± 0.005	0.901 ± 0.006	0.887 ± 0.006	0.874 ± 0.006	0.860 ± 0.006	0.856 ± 0.006
T-GRU-raw+s	0.897 ± 0.005	**0.925 ± 0.005**	0.900 ± 0.005	0.882 ± 0.006	0.871 ± 0.005	0.863 ± 0.005
T-GRU	0.893 ± 0.005	0.902 ± 0.006	0.898 ± 0.006	0.881 ± 0.006	0.868 ± 0.006	0.861 ± 0.006
T-GRU+u	**0.913 ± 0.005**	**0.927 ± 0.005**	**0.910 ± 0.005**	**0.897 ± 0.005**	**0.885 ± 0.005**	**0.878 ± 0.005**
DATA-GRU	**0.919 ± 0.004**	**0.934 ± 0.005**	**0.921 ± 0.005**	**0.907 ± 0.004**	**0.898 ± 0.005**	**0.893 ± 0.004**

used to adjust weights assigned to $x^s_{t\,deep}$. The expressions are as follows:

$$c^s_{t\,0/1} = F_{ARPF}(c_t) = \lfloor c_t - 0.5 \rfloor \tag{10}$$

$$x^s_{t\,0/true} = x_t \odot c^s_{t\,0/1} \tag{11}$$

$$x^s_{t\,deep} = TGRU(x^s_{t\,0/true}, w_{\Delta t}) \tag{12}$$

$$\alpha^s_{t\,deep} = TGRU(\alpha^s_{t\,0/1}, w_{\Delta t}) \tag{13}$$

$$x^s_t = x^s_{t\,deep} \odot \alpha^s_{t\,deep} \tag{14}$$

Interpretable embedding. To utilize information from both views, we combine them via an embedding layer. We select rectified linear unit (ReLU) as the activation function because it enables the learned representations to be interpretable. The expression is given below:

$$x^{adjust}_t = ReLU(W_emb[x^u_t; x^s_t] + b^{emb}) \tag{15}$$

The adjusted previous hidden status as given in Eq. (6) and the adjusted input as given in Eq. (15) are then injected into a standard GRU:

$$z_t = \delta(W_z x^{adjust}_t + U_z h^d_{t-1} + b_z) \tag{16}$$

$$r_t = \delta(W_r x^{adjust}_t + U_r h^d_{t-1} + b_r) \tag{17}$$

$$\tilde{h_t} = tanh(W x^{adjust}_t + U(r_t \odot h^d_{t-1}) + b) \tag{18}$$

$$h_t = (1 - z_t) \odot h^d_{t-1} + z_t \odot \tilde{h_t} \tag{19}$$

where $\delta(\bullet)$ is sigmoid function; W_z, U_z, W_r, U_r, W and U are trainable matrices; b_z, b_r and b are trainable vectors.

Objective Function. We use a softmax layer to generate the mortality risk scores from the hidden status at the last timestamp of the observation window:

$$\tilde{y_t} = softmax(W_{pred} h_t + b_{pred}) \tag{20}$$

where W_{pred} and b_{pred} are trainable matrix and vector.

We use cross-entropy as the objective function to calculate the classification loss between the true mortality label $\tilde{y}$ and the predicted label $\tilde{y_t}$ for each patient:

$$Loss(\tilde{y}, \tilde{y_t}) = \frac{1}{T_k} \sum_{t=1}^{T_k} (\tilde{y} \log \tilde{y_t} + (1 - \tilde{y}) \log(1 - \tilde{y_t})) \tag{21}$$

During the training process, the losses for all the patients in each minibatch are summed up to obtain the total loss for back propagation.

Experiments

Data Description and Experimental Settings

We conduct experiments on two real-world datasets, i.e., MIMIC-III (Johnson et al. 2016) and eICU Collaborative Research Dataset (Pollard et al. 2018).

MIMIC-III consists of medical records of 58K patients collected at Beth Israel Deaconess Medical Center over 11 years. We use the 20 most frequent laboratory parameters and the 30 most frequent chartevents as inputs. We set the observation window as 5 days and utilize all the patients meeting the following three conditions: 1) adult (aged 18 years or above); 2) at least one of the 20 laboratory parameters is not empty; 3) at least one of the 30 chartevents is not empty. We finally get a cohort of 34,660 patients. We conduct two kinds of prediction tasks: in-hospital mortality risk prediction (predict the likelihood of death for a patient during the treatment in hospital) and short-term mortality risk prediction (predict the likelihood of death for a patient a few days after the end of the observation window).

eICU consists of medical records of 200,859 patients collected from 208 critical care units in the United States between 2014 and 2015. We utilize the 50 most frequent laboratory parameters as inputs. We conduct in-hospital mortality risk prediction with different lengths of observation window (increase from the first day to the first 6 days after admission). We utilize all the adult patients with at least 2 visits within the observation window.

For both datasets, 70% of patients are randomly chosen as the training set and the rest patients are used as the test set. The experimental results are evaluated in terms of the area under the receiver operator characteristic curves (AUC).

Comparing Methods

- **Logistic Regression (LR) and Random Forests (RF):** As baselines, LR (Hosmer Jr, Lemeshow, and Sturdivant 2013) and RF (Breiman 2001) are used to model means of IMTS data since they are unable to model variable length sequences.
- **Independent Recurrent Neural Network (IndRNN):** Different from standard RNNs, neurons in IndRNN are

Table 2: The AUC scores ($mean \pm std$) of in-hospital mortality risk prediction with different observation windows for eICU. **Red** represents the best performance while **Blue** and **Green** indicate the second and third best performance, respectively.

Models	1 day	2 days	3 days	4 days	5 days	6 days
LR (Hosmer et al. 2013)	0.738 ± 0.003	0.768 ± 0.003	0.781 ± 0.003	0.792 ± 0.003	0.800 ± 0.003	0.801 ± 0.003
RF (Breiman 2001)	0.775 ± 0.004	0.792 ± 0.004	0.801 ± 0.003	0.813 ± 0.003	0.823 ± 0.003	0.820 ± 0.003
IndRNN (Li et al. 2018)	0.820 ± 0.003	0.838 ± 0.003	0.852 ± 0.003	0.862 ± 0.003	0.867 ± 0.003	0.877 ± 0.002
GRU-raw (Chung et al. 2014)	0.756 ± 0.004	0.825 ± 0.003	0.845 ± 0.003	0.857 ± 0.003	0.867 ± 0.003	0.875 ± 0.003
GRU (Chung et al. 2014)	0.816 ± 0.003	0.836 ± 0.003	0.852 ± 0.003	0.866 ± 0.003	0.869 ± 0.003	0.875 ± 0.002
T-LSTM (Baytas et al. 2017)	0.793 ± 0.003	0.819 ± 0.003	0.836 ± 0.003	0.845 ± 0.003	0.860 ± 0.003	0.869 ± 0.003
T-LSTM-ND (Baytas et al. 2017)	0.820 ± 0.004	0.839 ± 0.004	0.854 ± 0.004	0.867 ± 0.004	0.874 ± 0.004	0.879 ± 0.004
GRU-D (Che et al. 2018)	0.737 ± 0.004	0.803 ± 0.004	0.827 ± 0.003	0.849 ± 0.003	0.858 ± 0.003	0.861 ± 0.003
InterpNet (Shukla and Marlin 2019)	0.720 ± 0.004	0.805 ± 0.004	0.834 ± 0.003	0.847 ± 0.003	0.854 ± 0.003	0.857 ± 0.003
T-GRU-raw	0.768 ± 0.004	0.830 ± 0.003	0.847 ± 0.003	0.861 ± 0.003	0.870 ± 0.003	0.876 ± 0.002
T-GRU-raw+s	0.777 ± 0.004	0.832 ± 0.003	0.851 ± 0.003	0.863 ± 0.003	0.870 ± 0.003	0.880 ± 0.002
T-GRU	**0.822 ± 0.003**	**0.844 ± 0.003**	**0.859 ± 0.003**	**0.870 ± 0.003**	**0.875 ± 0.003**	**0.881 ± 0.002**
T-GRU+u	**0.826 ± 0.003**	**0.854 ± 0.003**	**0.866 ± 0.003**	**0.877 ± 0.002**	**0.883 ± 0.003**	**0.891 ± 0.002**
DATA-GRU	**0.836 ± 0.003**	**0.859 ± 0.003**	**0.872 ± 0.003**	**0.884 ± 0.002**	**0.890 ± 0.002**	**0.896 ± 0.002**

independent of each other in the same layer but are connected across different layers (Li et al. 2018).

- **T-LSTM and T-LSTM-ND:** T-LSTM (Baytas et al. 2017) described in the introduction section. Since T-LSTM is designed for longitudinal records, it decomposes data into long and short memory, which may not be suitable for ICU data. So we also compare with T-LSTM without such decomposition structure (T-LSTM-ND).
- **GRU-D:** GRU-D (Che et al. 2018) described in the introduction section.
- **InterpNet:** InterpNet (Shukla and Marlin 2019) described in the introduction section.
- **DATA-GRU variants:** Besides above methods, we consider six variants of DATA-GRU to verify the effectiveness of each component: (a) standard GRU for original records (GRU-raw) and augmented records (GRU); (b) T-GRU and T-GRU-raw are GRU and GRU-raw with the time-aware mechanism; (c) T-GRU-raw+s is T-GRU-raw with symptom-aware attention; (d) T-GRU+u is T-GRU with unreliability-aware attention.

Results and Discussion

The AUC scores of DATA-GRU and comparing methods for MIMIC-III and eICU are provided in Table 1 and Table 2. It can be seen that the AUC scores of DATA-GRU are continually larger than that of other methods for both datasets. E.g., for the in-hospital mortality prediction of MIMIC-III, the AUC score of DATA-GRU is 0.919, which is significantly larger than 0.903 achieved by InterpNet (Shukla and Marlin 2019). These results demonstrate that DATA-GRU achieves the best performance regardless of the prediction levels and the lengths of observation window. There are several possible reasons. Firstly, DATA-GRU introduces a time-aware structure to directly handle irregular intervals without the necessity of processing IMTS into equally-spaced, thus successfully preserving the contained useful information on the dynamics of patients' health status. Secondly, DATA-GRU designs novel dual-attention consisting of two novel attention mechanisms to handle missing values, thus capturing useful information from different views. Furthermore, all the parameters of DATA-GRU are optimized jointly, ensuring different components collaborate well to achieve global optimal. Besides, several other observations are drawn:

Firstly, DATA-GRU outperforms InterpNet (Shukla and Marlin 2019) by a large margin. InterpNet is the representative of processing IMTS into equally spaced, which may damage the medical considerations behind the sampling characteristic of EHR data and do not perform well when the number of original records is small, e.g., in-hospital mortality prediction for eICU with one day's observation data. In comparison, DATA-GRU avoids processing IMTS into equally spaced by introducing a time-aware structure to handle irregular intervals, achieving better results.

Secondly, DATA-GRU outperforms GRU-D (Che et al. 2018) and T-LSTM (including T-LSTM-ND, Baytas et al. 2017) significantly. This is because GRU-D and T-LSTM are unable to identify the differences in the reliability of different data points and thus cannot adjust the assigned weights. As a result, imputed values could play equal contributions with actual records even when they deviate from reality, thus severely affecting final prediction results.

Thirdly, variations of RNN with the time-aware mechanism generally outperform variations without such mechanism. E.g., for MIMIC-III, at the 1-day prediction level, AUC scores of T-GRU and T-LSTM-ND are 0.902 and 0.899, which are larger than 0.894 and 0.892 achieved by IndRNN and GRU. These results demonstrate that the introduction of the time-aware mechanism promotes the capacity of handling irregularly spaced data. Experimental results show that T-LSTM does not perform well for these tasks, which is probably because its decomposition structure damages the short-term information in ICU data.

Finally, modeling augmented records usually results in better results than directly modeling original records, e.g., for all the prediction tasks in eICU, T-GRU outperforms T-GRU-raw and T-GRU+u surpasses T-GRU-raw+s. This is probably because original EHR data is highly complex with both missing values and varying intervals, due to which the contained medical knowledge is difficult to extract. Conversely, temporal dependencies contained in aug-

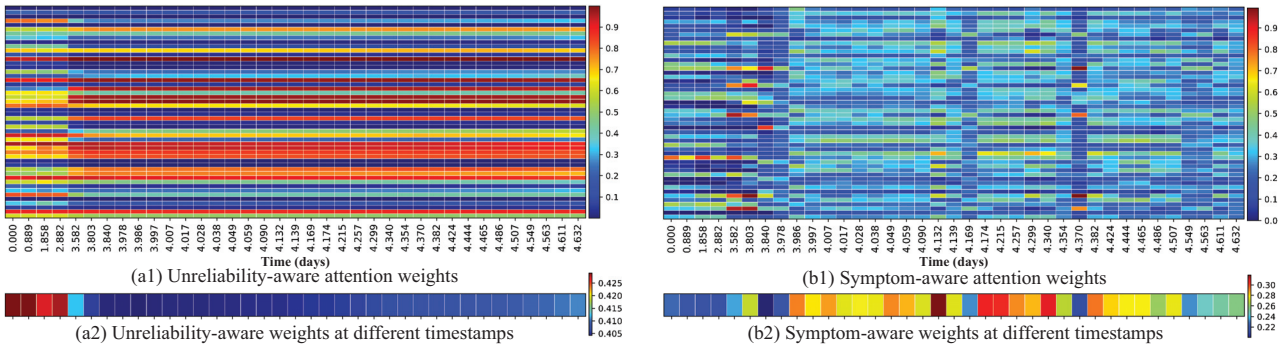


Figure 3: Visualization of attention weights. (a1) the weight matrix learned by unreliability-aware attention of DATA-GRU, which shows score assigned to each variable at each visit; (a2) the average weight of different variables at each visit; (b1) and (b2) provide weight matrix and average weight learned by symptom-aware attention of DATA-GRU.

mented data are easier to capture. In addition, T-GRU+u and T-GRU-raw+s outperform T-GRU and T-GRU-raw respectively, which demonstrate the effeteness of each attention mechanism. Furthermore, DATA-GRU consistently outperforms T-GRU+u and T-GRU-raw+s, which proves that the proposed dual-attention can effectively capture information from different views and improve final results.

Case Study

To demonstrate the benefit of applying DATA-GRU to real-world risk prediction tasks, we analyze attention weights learned by unreliability-aware and symptom-aware components of DATA-GRU. Fig. 3 shows a case study for predicting the in-hospital clinical outcome of a pneumonia patient in MIMIC-III based on the EHR data within 5 days after admission. The patient information is provided in Table 3. In Fig. 3, X-axis denotes time steps and Y-axis represents the learned attention score. Fig. 3(a1) shows the attention weight matrix learned by the unreliability-aware attention mechanism, which demonstrates that different attention scores are assigned to different elements in different time series. The average attention score of different variables at each visit is given in Fig. 3(a2), which shows that the unreliability-aware mechanism assigns larger weights on the first several records. This is because many physiological variables of the patient are actually examined in these visits. Furthermore, the unreliability-aware attention mechanism can effectively identify important variables. Fig. 4 shows attention weights contributed by all the 50 input variables. We can see that attention scores assigned to these variables are different. The item IDs of the 5 variables with the largest weights for the pneumonia patient are 87 (Braden Score), 787 (Carbon Dioxide), 742 (calprevflg), 646 (SpO2), and 51277 (RDW), most of which have been proved to be closely related to pneumonia (Sin, Man, and Marrie 2005). This proves that the unreliability-aware attention can effectively identify important variables and assign larger weights to reliable records to help them play important roles.

Fig. 3(b1) shows the attention matrix learned by the symptom-aware attention mechanism, which can be seen is different from Fig. 3(a1). This is because the unreliability-aware attention considers the quality of data, while the symptom-aware attention focuses on learning medical knowledge from original clinical records. The average attention score contributed by each visit is provided in Fig. 3(b2), which clearly shows that the symptom-aware attention focuses on visits in the middle. This is probably because the time intervals in the middle are small, due to which the symptoms of the patient change frequently within a certain period. All these results prove that the proposed dual-attention mechanism can analyze EHR data from different views and capture different aspects of information, thus effectively improving risk prediction results.

Table 3: Patient information in the case study. All the timestamps use admission time of the patient as the benchmark.

Diagnosis	In-hospital clinical outcome	Death time	Risk predicted by DATA-GRU
Pneumonia	Death	17.99 days	0.906

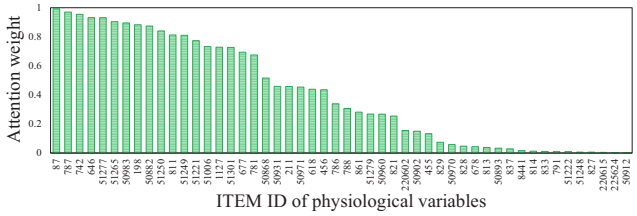


Figure 4: Attention weights assigned to different variables.

Conclusion

This paper proposes a novel DATA-GRU to predict the mortality risk of patients by using IMTS data. A time-aware mechanism is introduced to directly handle the irregular time intervals and preserve the contained useful information. Furthermore, a novel dual-attention mechanism is designed to tackle missing values in IMTS from both the data-quality view and the medical-knowledge view. Extensive experimental results and case study demonstrate that DATA-GRU outperforms existing methods significantly and provides in-

terpretable prediction results.

References

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE TPAMI*.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware lstm networks. In *KDD*, 65–74.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: natural language inference with natural language explanations. In *NIPS*, 9539–9549.

Che, C.; Xiao, C.; Liang, J.; Jin, B.; Zho, J.; and Wang, F. 2017. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *SIAM on Data Mining*, 198–206.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8(1):6085.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, 3504–3512.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. Gram: graph-based attention model for healthcare representation learning. In *KDD*, 787–795.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Comon, P. 2014. Tensors: a brief introduction. *IEEE Signal Processing Magazine* 31(3):44–53.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *CVPR*, 3146–3154.

Gao, L.; Li, X.; Song, J.; and Shen, H. T. 2019. Hierarchical lstms with adaptive attention for visual captioning. *IEEE TPAMI*.

Heo, J.; Lee, H. B.; Kim, S.; Lee, J.; Kim, K. J.; Yang, E.; and Hwang, S. J. 2018. Uncertainty-aware attention for reliable interpretation and prediction. In *NIPS*, 909–918.

Hosmer Jr, D. W.; Lemeshow, S.; and Sturdivant, R. X. 2013. *Applied logistic regression*, volume 398. John Wiley & Sons.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3:160035.

Kang, A.; Tan, Q.; Yuan, X.; Lei, X.; and Yuan, Y. 2017. Short-term wind speed prediction using eemd-lssvm model. *Advances in Meteorology* 2017:1–22.

Li, S.; Li, W.; Cook, C.; Zhu, C.; and Gao, Y. 2018. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *CVPR*, 5457–5466.

Lipton, Z. C.; Kale, D.; and Wetzel, R. 2016. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *Machine Learning for Healthcare Conference*, 253–270.

Liu, B.; Li, Y.; Sun, Z.; Ghosh, S.; and Ng, K. 2018a. Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In *AAAI*.

Liu, L.; Shen, J.; Zhang, M.; Wang, Z.; and Tang, J. 2018b. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. In *AAAI*.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*, 1903–1911.

Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, 743–752.

Pang, B.; Zha, K.; Cao, H.; Shi, C.; and Lu, C. 2019. Deep rnn framework for visual sequential applications. In *CVPR*, 423–432.

Pollard, T. J.; Johnson, A. E.; Raffa, J. D.; Celi, L. A.; Mark, R. G.; and Badawi, O. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data* 5.

Shankar, S., and Sarawagi, S. 2019. Posterior attention models for sequence to sequence learning. In *ICLR*.

Shickel, B.; Tighe, P. J.; Bihorac, A.; and Rashidi, P. 2017. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE JBHI* 22(5):1589–1604.

Shukla, S. N., and Marlin, B. M. 2019. Interpolation-prediction networks for irregularly sampled time series. In *ICLR*.

Sin, D. D.; Man, S. P.; and Marrie, T. J. 2005. Arterial carbon dioxide tension on admission as a marker of in-hospital mortality in community-acquired pneumonia. *The American journal of medicine* 118(2):145–150.

Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*.

Suo, Q.; Ma, F.; Canino, G.; Gao, J.; Zhang, A.; Veltri, P.; and Agostino, G. 2017. A multi-task framework for monitoring health conditions via attention-based recurrent neural networks. In *AMIA*, 1665–1674.

Tan, Q.; Ma, A. J.; Deng, H.; Wong, V. W.-S.; Tse, Y.-K.; Yip, T. C.-F.; Wong, G. L.-H.; Ching, J. Y.-L.; Chan, F. K.-L.; and Yuen, P.-C. 2018. A hybrid residual network and long short-term memory method for peptic ulcer bleeding mortality prediction. In *AMIA*, 998–1007.

Tan, Q.; Ma, A. J.; Ye, M.; Yang, B.; Deng, H.; Wong, V. W.-S.; Tse, Y.-K.; Yip, T. C.-F.; Wong, G. L.-H.; Ching, J. Y.-L.; et al. 2019. Ua-crnn: Uncertainty-aware convolutional recurrent neural network for mortality risk prediction. In *CIKM*, 109–118.

Xu, Y.; Biswal, S.; Deshpande, S. R.; Maher, K. O.; and Sun, J. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*, 2565–2573.

Ye, M.; Lan, X.; Wang, Z.; and Yuen, P. C. 2019a. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS*.

Ye, M.; Li, J.; Ma, A. J.; Zheng, L.; and Yuen, P. C. 2019b. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE TIP* 28(6):2976–2990.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019c. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 6210–6219.

Yuan, X.; Chen, C.; Yuan, Y.; Huang, Y.; and Tan, Q. 2015. Short-term wind power prediction based on lssvm–gsa model. *Energy Conversion and Management* 101:393–401.

Yuan, X.; Tan, Q.; Lei, X.; Yuan, Y.; and Wu, X. 2017. Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine. *Energy* 129:122–137.