Cooperative Joint Attentive Network for Patient Outcome Prediction on Irregular Multi-Rate Multivariate Health Data

Qingxiong Tan[1], **Mang Ye**[2*], **Grace Lai-Hung Wong**[3] and **PongChi Yuen**[1]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]School of Computer Science, Wuhan University, Wuhan, China
[3]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong
{csqxtan, pcyuen}@comp.hkbu.edu.hk, yemang@whu.edu.cn, wonglaihung@cuhk.edu.hk

Abstract

Due to the dynamic health status of patients and discrepant stability of physiological variables, health data often presents as irregular multi-rate multivariate time series (IMR-MTS) with significantly varying sampling rates. Existing methods mainly study changes of IMR-MTS values in the time domain, without considering their different dominant frequencies and varying data quality. Hence, we propose a novel Cooperative Joint Attentive Network (CJANet) to analyze IMR-MTS in frequency domain, which adaptively handling discrepant dominant frequencies while tackling diverse data qualities caused by irregular sampling. In particular, novel dual-channel joint attention is designed to jointly identify important magnitude and phase signals while detecting their dominant frequencies, automatically enlarging the positive influence of key variables and frequencies. Furthermore, a new cooperative learning module is introduced to enhance information exchange between magnitude and phase channels, effectively integrating global signals to optimize the network. A frequency-aware fusion strategy is finally designed to aggregate the learned features. Extensive experimental results on real-world medical datasets indicate that CJANet significantly outperforms existing methods and provides highly interpretable results.

1 Introduction

The rapid growth of electronic health records (EHR) provides good chances to build models to improve healthcare quality. One important task is to predict the mortality risk of patients based on their historical records, which can identify high-risk patients. This task is challenging because of the diverse and changing sampling rates (termed as irregular multi-rate) in different variables, as illustrated in Fig. 1. Different variables have diverse sampling rates, which reflect their discrepant stabilities and fluctuation frequencies. Furthermore, the sampling rates within each sequence vary significantly because of the dynamic changes in the health conditions of patients.

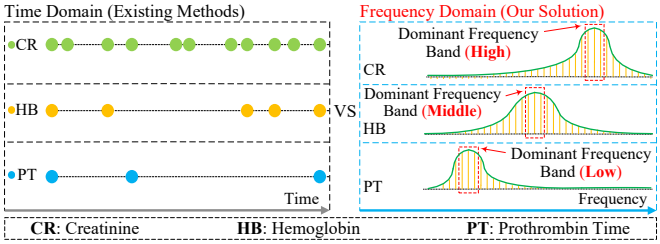


Figure 1: Advantages of modeling IMR-MTS data in the frequency domain. Since variables have discrepant stability, they are examined under diverse rates, e.g., low for PT while high for CR. Time-domain methods mainly study changes of values over time, which may not extract vital medical clues in sampling rates. In contrast, we observe that dominant frequencies of variables with different stability locate in diverse bands, providing feasible solutions for mining such clues.

Since standard models are designed for data with equal intervals [Pang *et al.*, 2020; Ye *et al.*, 2020; Pang *et al.*, 2021], existing methods often process irregular records into equally-spaced data by discretizing time axis into non-overlapping intervals with a constant interval, e.g., 1 hour for ICU data [Xu *et al.*, 2018], 6 months for chronic disease [Tan *et al.*, 2018], or confirmed by learning [Shukla and Marlin, 2019]. Missing values are then handled via imputation techniques, e.g., alignment methods [Tan *et al.*, 2018] or interpolation approaches [Che *et al.*, 2018a]. There are two major limitations. First, these methods are unable to handle discrepant data quality caused by varying sampling rates. As a result, low-quality data (e.g., imputed or interpolated records with much noise) may play important roles, which impair their results. Second, they ignore the underlying vital medical impacts reflected by the pattern of varying sampling rates. Specifically, highly changeable variables are usually examined densely to timely monitor the health status of patients, while stable variables are measured sparsely. Hence, variables with diverse sampling rates often have discrepant dominant frequencies and function at different bands, as illustrated in Fig. 1. However, these methods analyze medical records in the *time domain*, which mainly study the changes of variable values over time. As a result, the important medical clues in the varying sampling rates are ignored, which may limit their performance.

To address the aforementioned challenges, we propose a novel Cooperative Joint Attentive Network (CJANet) for

*Corresponding Author

medical IMR-MTS analysis in the *frequency domain*. To effectively integrate frequency magnitudes and phases while adaptively tackling the issues of discrepant data quality and dominant frequency, we propose a novel dual-channel joint attention mechanism to dynamically learn contribution scores of both features at different frequencies. In particular, dual-channel joint attention weights are learned to simultaneously adjust contributions of magnitude and phase signals at diverse frequency bands, globally optimized in the end-to-end training network. This strategy effectively identifies vital variables and frequencies, automatically enlarging their positive influence to provide accurate interpretable prediction results. Furthermore, at each frequency point, magnitude and phase have corresponding relationships, whose combinations at different frequencies can restore the original temporal signal. In view of this, we design a new cooperative learning module with dual channels to incorporate magnitudes and phases while enhancing the information-sharing capabilities between these channels. This module incorporates an information exchange structure for both channels to timely share their learned knowledge, effectively utilizing global signals to optimize parameters of the network and cooperate closely to achieve accurate results. Finally, we introduce a frequency-aware fusion structure to learn aggregation weights to identify key frequencies and enhance their contributions.

The main contributions of this paper are listed as follows:

- We start the first attempt to model the IMR-MTS data in the frequency domain by proposing a novel CJANet to jointly deal with the discrepant dominant frequencies and diverse data quality problems, which effectively tackle the irregular sampling rates in health records.
- We design a new dual-channel joint attention mechanism to dynamically adjust the contributions of frequency magnitudes and phases at diverse frequencies, thus automatically strengthening the positive influence of key variables and dominant frequencies.
- We propose a cooperative learning module into the deep learning architecture to simultaneously model the magnitude and phase signals while enhancing their information-sharing capabilities, effectively utilizing global information to optimize their parameters.
- We empirically demonstrate that CJANet outperforms the state-of-the-art methods on real-world medical datasets. The case study indicates that the obtained clinical risk prediction results are highly interpretable.

2 Related Work

Attention Mechanism for Health Data. [Choi *et al.*, 2016] built RETAIN to identify vital visits and features using attention. [Choi *et al.*, 2017] introduced attention to learn robust representations of health data. [Xu *et al.*, 2018] built RAIM with efficient attention to jointly handle continuous and discrete data. [Tan *et al.*, 2020b] designed uncertainty-aware attention to achieve explainable predictions. [Tan *et al.*, 2020a] built DATA-GRU to jointly handle missing values and varying time intervals via dual-attention and time-aware mechanisms. Attention is also applied to the noisy problem [Heo *et al.*, 2018] and clinical context embedding [Qiao *et al.*, 2018]. These mechanisms improve the performance and interpretability of networks at some extent [Tan *et al.*, 2021]. However, these methods analyze health data in the time domain, which cannot be applied to the irregular EHR data consisting of multiple physiological variables with very different sampling rates and fluctuation frequencies.

Frequency Analysis for Health Data. Several works have applied frequency analysis for health data [Fang *et al.*, 2020]. [Parhi and Zhang, 2019] designed a frequency-domain model ratio method to select spectral power features for seizure prediction. [Issa *et al.*, 2020] proposed a novel user-independent method to classify emotion using the electroencephalograph (EEG) brain signals. [Zhang *et al.*, 2020] built a noninvasive system to monitor blood glucose by using a fitting-based sliding window algorithm to analyze smartphone photoplethysmography (PPG) signals. However, these models are usually designed for regular signals, e.g., EEG and PPG, which is unsuitable for the irregular EHR data. Furthermore, these methods assume that different data have equal quality, which cannot deal with the discrepant data quality problem.

3 Proposed Method

We propose a novel Cooperative Joint Attentive Network (CJANet) for IMR-MTS in the frequency domain, as shown in Fig. 2. It contains three parts, namely, dual-channel joint attention, cooperative learning, and frequency-aware fusion. Specifically, joint attention dynamically learns contribution weights of magnitudes and phases while detecting vital frequency bands, enlarging the influence of key factors to provide accurate and interpretable predictions. Cooperative learning incorporates magnitudes and phases while introducing an information-exchange unit for them to conveniently share their learned knowledge, integrating global information to promote results. A frequency-aware fusion structure is designed to adaptively learn aggregation weights to further adjust the contributions of features at diverse frequencies, effectively strengthening the positive impact of important features.

3.1 Dual-channel Joint Attention in Frequency Domain

Let $\boldsymbol{X}^{ir} = [\boldsymbol{x}_1^{ir}, \cdots, \boldsymbol{x}_n^{ir}, \cdots, \boldsymbol{x}_N^{ir}]$ represent a tuple with N irregular time series and $\boldsymbol{T}^{ir} = [\boldsymbol{t}_1^{ir}, \cdots, \boldsymbol{t}_n^{ir}, \cdots, \boldsymbol{t}_N^{ir}]$ denote the tuple of corresponding examination timestamps. For IMR-MTS data, the challenge manifests in two aspects: 1) different variables have diverse sampling rates; 2) even within the sequence of each variable, its sampling rates vary significantly, as illustrated in Fig. 1. To obtain regular data, which can serve as input of arbitrary standard machine learning models, the mean function of Gaussian process (GP) [Zhang and Williamson, 2019] m_{GP}, is adopted to generate regular data $\boldsymbol{x}_n$ by inputting equally-spaced timestamps $\boldsymbol{t}_n$ based on observed variable values $\boldsymbol{x}_n^{ir}$ and examination timestamps $\boldsymbol{t}_n^{ir}$:

$$\boldsymbol{x}_n = m_{\mathrm{GP}}(\boldsymbol{t}_n | \boldsymbol{x}_n^{ir}, \boldsymbol{t}_n^{ir}). \tag{1}$$

Because of the irregular sampling rates, data in $\boldsymbol{x}_n$ may have diverse quality. Confidence interval (*CI*) can handle this issue because it describes the degree in which the estimated

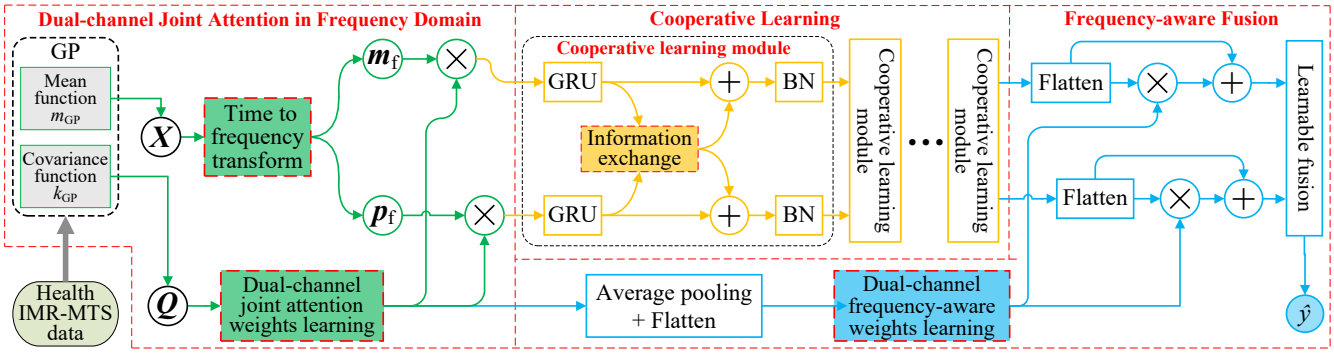


Figure 2: The overall framework of the proposed method.

value deviates from the true value [Zhang and Williamson, 2019], e.g., large *CI* values reflect a huge difference between estimated and true values, which indicate that estimated data may contain much noise and is of low quality. Thus, *CI* values $\boldsymbol{v}_n$ is calculated via the covariance function of GP k_{GP}:

$$\boldsymbol{v}_n = z_{CI} \times k_{\text{GP}}(\boldsymbol{t}_n|\boldsymbol{t}_n^{ir}), \tag{2}$$

where z_{CI} is a scaling factor decided by the level of *CI*, which is selected as 95% in our implementation.

Let $\boldsymbol{X}$ denote the generated regular sequences, i.e., $\boldsymbol{X} = [\boldsymbol{x}_0, \cdots, \boldsymbol{x}_n, \cdots, \boldsymbol{x}_{N-1}] \in \mathcal{R}^{L\times N}$, where N is the number of variables; $\boldsymbol{x}_n = [x_{0,n}; \cdots; x_{l,n}; \cdots; x_{L-1,n}]$, where L is the length of the generated sequence. Simultaneously, $\boldsymbol{v}_n$ is converted to quality score via $\boldsymbol{q}_n = (\boldsymbol{v}_{max} - \boldsymbol{v}_n)/(\boldsymbol{v}_{max} - \boldsymbol{v}_{min})$, where $\boldsymbol{v}_{max}$ and $\boldsymbol{v}_{min}$ are the maximum and minimum values in $\boldsymbol{v}_n$. Thus, we obtain a matrix $\boldsymbol{Q}$ of quality scores, which quantitatively describe the reliability of every element in matrix $\boldsymbol{X}$, represented as $\boldsymbol{Q} = [\boldsymbol{q}_0, \cdots, \boldsymbol{q}_n, \cdots, \boldsymbol{q}_{N-1}] \in \mathcal{R}^{L\times N}$, where $\boldsymbol{q}_n = [q_{0,n}; \cdots; q_{l,n}; \cdots; q_{L-1,n}]$. Frequency transformation [Lin and others, 2016] is adopted to convert sequences in $\boldsymbol{X}$ to the frequency domain and represent it as the combination of real part $\text{Re}(z_{\text{f},n})$ and imaginary part $\text{Im}(z_{\text{f},n}) * i$:

$$\begin{aligned} z_{\text{f},n} &= \sum_{l=0}^{L-1} x_{l,n} w_L^{\text{fl}}, \text{f} = 0, 1, \cdots, F-1, \\ &= \text{Re}(z_{\text{f},n}) + \text{Im}(z_{\text{f},n}) * i, \end{aligned} \tag{3}$$

where $w_L = e^{-i\frac{2\pi}{L}}$; $i = \sqrt{-1}$ is the unit imaginary root; F is the length of the spectrum.

We calculate magnitude $m_{\text{f},n} = \sqrt{\text{Re}(z_{\text{f},n})^2 + \text{Im}(z_{\text{f},n})^2}$ and phase $p_{\text{f},n} = \tan^{-1}(\text{Im}(z_{\text{f},n})/\text{Re}(z_{\text{f},n}))$. The frequency magnitudes of different variables are thus represented as $\boldsymbol{M} = [\boldsymbol{m}_0, \cdots, \boldsymbol{m}_n, \cdots, \boldsymbol{m}_{N-1}] \in \mathcal{R}^{F\times N}$, where $\boldsymbol{m}_n = [m_{0,n}; \cdots; m_{\text{f},n}; \cdots; m_{F-1,n}]$, and phase signals are denoted as $\boldsymbol{P} = [\boldsymbol{p}_0, \cdots, \boldsymbol{p}_n, \cdots, \boldsymbol{p}_{N-1}] \in \mathcal{R}^{F\times N}$, where $\boldsymbol{p}_n = [p_{0,n}; \cdots; p_{\text{f},n}; \cdots; p_{F-1,n}]$. Batch normalization (*BN*) is adopted to enhance the compatibility of both signals. Matrix $\boldsymbol{Q}$ can describe the quality of different elements in each sequence but is unable to identify which sequence is more important for estimating the health status of patients. Several different variables are examined at same rates may simply because they are measured from the same fluid of patients. For example, both *Hematocrit* (*HCT*) and *Hemoglobin* (*HB*) are examined from blood and often recorded synchronously, but they may contribute differently for estimating health conditions. Joint attention weights $\boldsymbol{\alpha}_{j,m}$ are thus learned for magnitude features and batch normalized:

$$\boldsymbol{\alpha}_{j,m} = BN(\text{sigmoid}(\boldsymbol{W}_{j,m}^T \boldsymbol{Q} + \boldsymbol{b}_{j,m})), \tag{4}$$

where $\boldsymbol{W}_{j,m} \in \mathcal{R}^{L\times F}$ is a trainable weight vector; $\boldsymbol{b}_{j,m} \in \mathcal{R}^N$ is a trainable bias vector; *BN* is batch normalization.

Magnitude and phase signals may not be synchronized in the changing patterns. Even for same variable, its magnitude and phase often fluctuate at different frequencies. Therefore, to enable both signals to function at their own paces, instead of using a weights-sharing attention module, we learn another attention weights $\boldsymbol{\alpha}_{j,p}$ for phases:

$$\boldsymbol{\alpha}_{j,p} = BN(\text{sigmoid}(\boldsymbol{W}_{j,p}^T \boldsymbol{Q} + \boldsymbol{b}_{j,p})), \tag{5}$$

where $\boldsymbol{W}_{j,p} \in \mathcal{R}^{L\times F}$ and $\boldsymbol{b}_{j,p} \in \mathcal{R}^N$ are trainable weight and bias vectors, respectively.

The learned weights are used to adjust the contributions of magnitude $\boldsymbol{M}$ and phase $\boldsymbol{P}$:

$$\boldsymbol{M}^j = \boldsymbol{M} \odot \boldsymbol{\alpha}_{j,m}, \tag{6}$$

$$\boldsymbol{P}^j = \boldsymbol{P} \odot \boldsymbol{\alpha}_{j,p}, \tag{7}$$

where $\odot$ represents Hadamard product.

3.2 Cooperative Learning

Since the changes of magnitudes and phases along frequency bands contain vital sequential information, we adopt RNNs rather than non-sequential models to build the cooperative learning module. Specifically, Gated Recurrent Unit (GRU) [Chung *et al.*, 2014] is adopted, which can be replaced with other RNN variants by simple modifications. The adjusted magnitudes $\boldsymbol{M}^j$ are modeled along different frequencies:

$$\boldsymbol{h}_{\text{f}}^m = \text{GRU}(\boldsymbol{m}_{f,:}, \boldsymbol{h}_{\text{f}-1}^m), \tag{8}$$

where $\boldsymbol{m}_{f,:} = [m_{f,0}^j, \cdots, m_{\text{f},n}^j, \cdots, m_{f,N-1}^j]$ is the elements of the fth frequency in $\boldsymbol{M}^j$; $\boldsymbol{h}_{\text{f}}^m \in \mathcal{R}^{d^m}$ is the hidden states learned for the magnitude signal and d^m is the number of hidden units in the magnitude-channel GRU network.

Models	In-hospital	10 days	20 days	30 days	40 days	50 days
RF [Breiman, 2001]	0.777 ± 0.007	0.781 ± 0.007	0.776 ± 0.007	0.768 ± 0.006	0.767 ± 0.006	0.756 ± 0.006
LR [Hosmer Jr *et al.*, 2013]	0.781 ± 0.008	0.780 ± 0.008	0.769 ± 0.008	0.766 ± 0.008	0.756 ± 0.007	0.750 ± 0.007
GRU [Chung *et al.*, 2014]	0.838 ± 0.006	0.804 ± 0.007	0.790 ± 0.006	0.783 ± 0.007	0.781 ± 0.006	0.779 ± 0.006
PLSTM [Neil *et al.*, 2016]	0.826 ± 0.006	0.835 ± 0.006	0.825 ± 0.006	0.797 ± 0.006	0.791 ± 0.006	0.779 ± 0.006
IndRNN [Li *et al.*, 2018]	0.819 ± 0.007	0.824 ± 0.007	0.818 ± 0.006	0.808 ± 0.006	0.805 ± 0.006	0.797 ± 0.006
GRU-D [Che *et al.*, 2018b]	0.884 ± 0.005	0.877 ± 0.005	0.864 ± 0.005	0.857 ± 0.005	0.848 ± 0.006	0.834 ± 0.005
InterpNet [Shukla and Marlin, 2019]	0.881 ± 0.005	0.856 ± 0.006	0.848 ± 0.005	0.838 ± 0.005	0.833 ± 0.005	0.832 ± 0.005
DATA-GRU [Tan *et al.*, 2020a]	**0.896 ± 0.005**	0.880 ± 0.005	0.874 ± 0.005	0.870 ± 0.005	**0.862 ± 0.005**	**0.858 ± 0.005**
MLSTM [Zhao *et al.*, 2020]	0.846 ± 0.006	0.832 ± 0.006	0.824 ± 0.006	0.823 ± 0.006	0.820 ± 0.005	0.820 ± 0.005
GRU-m	0.864 ± 0.005	0.863 ± 0.005	0.846 ± 0.005	0.842 ± 0.005	0.840 ± 0.005	0.827 ± 0.005
GRU-m+j	0.880 ± 0.005	0.872 ± 0.005	0.861 ± 0.005	0.850 ± 0.005	0.841 ± 0.005	0.830 ± 0.005
GRU-p	0.817 ± 0.006	0.790 ± 0.007	0.782 ± 0.006	0.773 ± 0.006	0.765 ± 0.006	0.753 ± 0.006
GRU-p+j	0.841 ± 0.006	0.833 ± 0.006	0.817 ± 0.006	0.804 ± 0.006	0.792 ± 0.006	0.784 ± 0.006
Twin-GRU	0.878 ± 0.005	0.871 ± 0.005	0.862 ± 0.005	0.848 ± 0.005	0.845 ± 0.005	0.843 ± 0.005
Twin-GRU+jc	0.891 ± 0.005	**0.896 ± 0.005**	**0.886 ± 0.005**	**0.874 ± 0.005**	**0.861 ± 0.005**	**0.858 ± 0.005**
Twin-GRU+jf	**0.895 ± 0.004**	**0.894 ± 0.005**	**0.884 ± 0.005**	**0.873 ± 0.004**	**0.862 ± 0.005**	**0.859 ± 0.005**
CJANet	**0.901 ± 0.004**	**0.898 ± 0.005**	**0.890 ± 0.004**	**0.877 ± 0.004**	**0.870 ± 0.004**	**0.864 ± 0.004**

Table 1: AUC score ($mean \pm std$) of risk prediction on MIMIC-III. **Red**, **Blue** and **Green** represent the best, the second and third best results.

Simultaneously, to extract the sequential changes of phase information along frequencies, a symmetric GRU neural network is constructed for $\boldsymbol{P}_j$ by modeling the adjusted phase signals over different frequency bands:

$$\boldsymbol{h}_{\mathrm{f}}^{p} = \mathrm{GRU}(\boldsymbol{p}_{f,:}, \boldsymbol{h}_{\mathrm{f}-1}^{p}), \tag{9}$$

where $\boldsymbol{p}_{f,:} = [p_{f,0}^{j}, \cdots, p_{\mathrm{f},n}^{j}, \cdots, p_{f,N-1}^{j}]$ is the elements of the fth frequency in $\boldsymbol{P}^{j}$; $\boldsymbol{h}_{\mathrm{f}}^{p} \in \mathcal{R}^{d^p}$ is the hidden states learned for the phase signal and d^p is the number of hidden units in the phase-channel network.

The hidden states learned in the magnitude-channel unit $\boldsymbol{h}_{\mathrm{f}}^{m}$ and the states learned in the phase-channel network $\boldsymbol{h}_{\mathrm{f}}^{p}$ are then integrated to learn cooperative information $\boldsymbol{h}^{co}$ via:

$$\boldsymbol{h}_{\mathrm{f}}^{co} = \boldsymbol{h}_{\mathrm{f}}^{m} \odot \boldsymbol{h}_{\mathrm{f}}^{p}. \tag{10}$$

The learned cooperative information $\boldsymbol{h}^{co}$ is further fed back to magnitude and phase channels respectively and simultaneously to adjust their status, thus informing them of the learning status of each other. *BN* is conducted for the dual-channel informed states to increase information compatibility. This learning strategy enables both channels to receive global information enabling better results with global optimization.

$$\boldsymbol{h}_{\mathrm{f}}^{m,adj} = BN(\boldsymbol{h}_{\mathrm{f}}^{m} + \boldsymbol{h}_{\mathrm{f}}^{co}), \tag{11}$$

$$\boldsymbol{h}_{\mathrm{f}}^{p,adj} = BN(\boldsymbol{h}_{\mathrm{f}}^{p} + \boldsymbol{h}_{\mathrm{f}}^{co}). \tag{12}$$

The learned states are then injected into deeper cooperative learning modules to repeat the above procedures so as to learn more informative and accurate representations. For convenience, the number of hidden units is set the same for magnitude and phase channels at each layer. Specifically, in our implementation, a three-layer cooperative learning structure is built with the hidden units of 32, 16, and 1, respectively.

3.3 Frequency-aware Fusion

Hidden states at all the frequencies of the last layer of the cooperative learning module are utilized to produce final results. States of magnitude and phase channels are flattened:

$$\boldsymbol{h}_{\mathrm{fla}}^{m} = [h_{0,l}^{m}, h_{1,l}^{m}, \ldots, h_{F-1,l}^{m}]^{T}, \tag{13}$$

$$\boldsymbol{h}_{\mathrm{fla}}^{p} = [h_{0,l}^{p}, h_{1,l}^{p}, \ldots, h_{F-1,l}^{p}]^{T}. \tag{14}$$

Frequency-aware weights are learned to aggregate features at different frequencies. Average pooling and flatten operations are performed for the 2^{nd} dimension of $\boldsymbol{\alpha}_{j,m}$ and $\boldsymbol{\alpha}_{j,p}$:

$$\boldsymbol{\alpha}_{af,m} = \mathrm{Flatten}(\mathrm{AvePooling}_{2^{nd}}(\boldsymbol{\alpha}_{j,m})), \tag{15}$$

$$\boldsymbol{\alpha}_{af,p} = \mathrm{Flatten}(\mathrm{AvePooling}_{2^{nd}}(\boldsymbol{\alpha}_{j,p})). \tag{16}$$

The frequency-aware fusion weights of magnitudes and phase are therefore learned and batch normalized:

$$\boldsymbol{\alpha}_{fuse,m} = BN(\mathrm{sigmoid}(\boldsymbol{W}_{fuse,m}^{T}\boldsymbol{\alpha}_{af,m} + \boldsymbol{b}_{fuse,m})), \tag{17}$$

$$\boldsymbol{\alpha}_{fuse,p} = BN(\mathrm{sigmoid}(\boldsymbol{W}_{fuse,p}^{T}\boldsymbol{\alpha}_{af,p} + \boldsymbol{b}_{fuse,p})), \tag{18}$$

where $\boldsymbol{W}_{fuse,m} \in \mathcal{R}^{F\times F}$ and $\boldsymbol{W}_{fuse,p} \in \mathcal{R}^{F\times F}$ are trainable weight vectors; $\boldsymbol{b}_{fuse,m} \in \mathcal{R}^{F}$ and $\boldsymbol{b}_{fuse,p} \in \mathcal{R}^{F}$ are trainable bias vectors.

The learned weights are applied to adjust contributions of $\boldsymbol{h}_{fla}^{m}$ and $\boldsymbol{h}_{fla}^{p}$ at different frequencies while a shortcut connection is adopted to facilitate information propagation:

$$\boldsymbol{h}_{fuse,rl}^{m} = \boldsymbol{\alpha}_{fuse,m} \odot \boldsymbol{h}_{fla}^{m} + \boldsymbol{h}_{fla}^{m}, \tag{19}$$

$$\boldsymbol{h}_{fuse,rl}^{p} = \boldsymbol{\alpha}_{fuse,p} \odot \boldsymbol{h}_{fla}^{p} + \boldsymbol{h}_{fla}^{p}. \tag{20}$$

With a shortcut connection, $\boldsymbol{h}_{fuse,rl}^{m}$ and $\boldsymbol{h}_{fuse,rl}^{p}$ are aggregated $\boldsymbol{h}^{m,p} = [\boldsymbol{h}_{fuse,rl}^{m}; \boldsymbol{h}_{fuse,rl}^{p}]$ while weights are learned to adjust contributions of different features:

$$\boldsymbol{\alpha}_{m,p} = BN(\mathrm{sigmoid}(\boldsymbol{W}_{m,p}^{T}\boldsymbol{h}^{m,p} + \boldsymbol{b}_{m,p})), \tag{21}$$

$$\boldsymbol{h}_{final} = \boldsymbol{\alpha}_{m,p} \odot \boldsymbol{h}^{m,p} + \boldsymbol{h}^{m,p}, \tag{22}$$

where $\boldsymbol{W}_{m,p} \in \mathcal{R}^{2F\times 2F}$ is a trainable weight vector; $\boldsymbol{b}_{m,p} \in \mathcal{R}^{2F}$ is a trainable bias vector.

A fully connected layer with the softmax activation is utilized to produce final risk scores from $\boldsymbol{h}_{final}$:

$$\tilde{y} = \mathrm{softmax}(\boldsymbol{W}_{dense}^{T}\boldsymbol{h}_{final} + \boldsymbol{b}_{dense}), \tag{23}$$

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

Models	4 months	8 months	12 months	16 months	20 months	24 months
RF [Breiman, 2001]	0.841 ± 0.014	0.834 ± 0.013	0.826 ± 0.013	0.818 ± 0.013	0.816 ± 0.013	0.813 ± 0.013
LR [Hosmer Jr *et al.*, 2013]	0.813 ± 0.013	0.812 ± 0.014	0.806 ± 0.013	0.803 ± 0.013	0.796 ± 0.014	0.792 ± 0.014
GRU [Chung *et al.*, 2014]	0.863 ± 0.008	0.851 ± 0.009	0.849 ± 0.009	0.846 ± 0.009	0.839 ± 0.009	0.836 ± 0.009
PLSTM [Neil *et al.*, 2016]	0.853 ± 0.009	0.848 ± 0.009	0.845 ± 0.009	0.842 ± 0.009	0.828 ± 0.009	0.824 ± 0.009
IndRNN [Li *et al.*, 2018]	0.862 ± 0.008	0.854 ± 0.009	0.848 ± 0.009	0.844 ± 0.009	0.843 ± 0.009	0.822 ± 0.010
GRU-D [Che *et al.*, 2018b]	0.926 ± 0.007	0.921 ± 0.007	**0.919 ± 0.007**	0.904 ± 0.008	**0.894 ± 0.008**	0.882 ± 0.008
InterpNet [Shukla and Marlin, 2019]	0.928 ± 0.006	0.922 ± 0.006	0.912 ± 0.007	0.902 ± 0.007	0.890 ± 0.007	**0.884 ± 0.008**
DATA-GRU [Tan *et al.*, 2020a]	0.922 ± 0.006	0.916 ± 0.007	0.909 ± 0.007	0.893 ± 0.007	0.886 ± 0.008	0.870 ± 0.008
MLSTM [Zhao *et al.*, 2020]	0.889 ± 0.008	0.875 ± 0.008	0.870 ± 0.008	0.867 ± 0.008	0.866 ± 0.008	0.840 ± 0.009
GRU-m	0.855 ± 0.009	0.852 ± 0.009	0.847 ± 0.009	0.834 ± 0.009	0.811 ± 0.010	0.792 ± 0.010
GRU-m+j	0.890 ± 0.008	0.863 ± 0.008	0.856 ± 0.008	0.843 ± 0.009	0.826 ± 0.009	0.822 ± 0.010
GRU-p	0.781 ± 0.011	0.766 ± 0.011	0.758 ± 0.011	0.754 ± 0.011	0.739 ± 0.011	0.729 ± 0.012
GRU-p+j	0.896 ± 0.008	0.832 ± 0.010	0.803 ± 0.010	0.797 ± 0.011	0.781 ± 0.010	0.748 ± 0.011
Twin-GRU	0.874 ± 0.008	0.860 ± 0.009	0.856 ± 0.009	0.849 ± 0.009	0.848 ± 0.009	0.846 ± 0.009
Twin-GRU+jc	**0.933 ± 0.006**	**0.925 ± 0.006**	0.915 ± 0.006	**0.905 ± 0.007**	0.885 ± 0.008	0.854 ± 0.008
Twin-GRU+jf	**0.936 ± 0.006**	**0.930 ± 0.006**	**0.928 ± 0.006**	**0.920 ± 0.006**	**0.908 ± 0.007**	**0.888 ± 0.007**
CJANet	**0.941 ± 0.006**	**0.936 ± 0.006**	**0.932 ± 0.006**	**0.925 ± 0.006**	**0.913 ± 0.007**	**0.895 ± 0.007**

Table 2: AUC score ($mean \pm std$) of risk prediction on PUB. **Red**, **Blue** and **Green** represent the best, the second and third best results.

4 Experiments

4.1 Data Description and Experimental Settings

Experiments are conducted on two EHR datasets. MIMIC-III [Johnson *et al.*, 2016] contains EHR of 58K patients at Beth Israel Deaconess Medical Center over 11 years. Thirty types of common lab test results in the first ten days are used. The cohort of all 38549 adults is used. We conduct in-hospital and short-term mortality risk predictions to evaluate the likelihood of death for a patient during the treatment in hospital and a few days after observation period.

The Peptic Ulcer Bleeding (PUB) dataset contains EHR data of 6367 patients at Prince of Wales Hospital over 10 years. We utilize seven types of popular lab test results as inputs. Patients in this dataset have long records, which usually last several years. Therefore, we perform long-term mortality risk predictions.

Implementation details. We randomly choose 70% of patients to train and use rest patients to test. The results are evaluated via the area under the receiver operator characteristic curves (AUC). We use cross-entropy as the loss function and Adam as the optimizer. The networks are trained for 30 epochs with a batch size of 32. The initial learning rate is set as 0.005 and decay by 10% every 3 epochs.

4.2 Comparing Methods

- **RF and LR:** RF [Breiman, 2001] and LR [Hosmer Jr *et al.*, 2013] are used baselines.
- **PLSTM:** [Neil *et al.*, 2016] converges faster than standard LSTM by using a parametrized oscillation to control time gates.
- **IndRNN:** To handle gradient vanishing issue, [Li *et al.*, 2018] connects neurons of different layers but makes neurons in the same layer independent.
- **GRU-D:** [Che *et al.*, 2018b] imputes missing data as decay of previously values toward empirical mean over time. Missing patterns are utilized to improve results.
- **InterpNet:** [Shukla and Marlin, 2019] trains an interpolation network to process irregular records while using a prediction network to model the obtained data.
- **DATA-GRU:** [Tan *et al.*, 2020a] designs dual-attention and time-aware mechanisms to deal with missing values and varying intervals.
- **MLSTM:** [Zhao *et al.*, 2020] introduces a memory filter structure controlled via a learnable parameter to increase the ability for long-term memory.
- **CJANet variants:** Besides above methods, we consider seven variants to verify the effectiveness of each component: (a) standard GRU for magnitudes (GRU-m) and phases (GRU-p); (b) GRU-m+j and GRU-p+j are GRU-m and GRU-p with joint attention; (c) Twin-GRU is two-channel GRU for magnitudes and phases; (d) Twin-GRU+jc is Twin-GRU with joint attention and cooperative learning mechanism; (e) Twin-GRU+jf is Twin-GRU with joint attention and frequency-aware fusion.

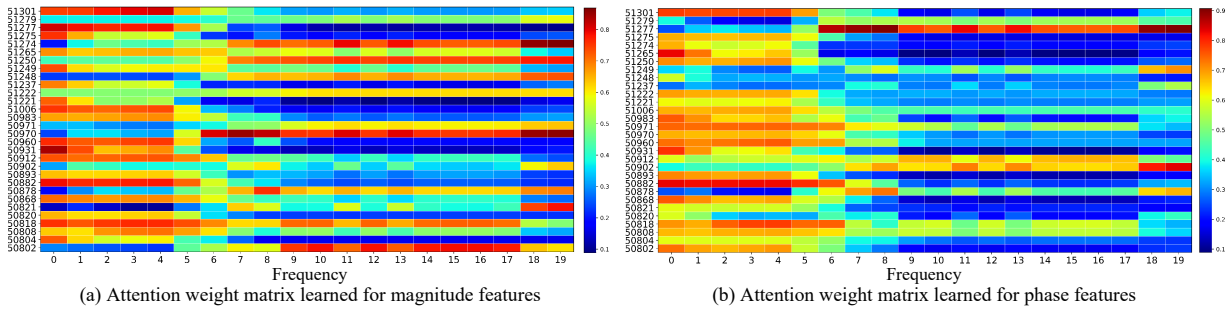
4.3 Results and Discussion

The experimental results on MIMIC-III and PUB are provided in Tables 1 and 2, which indicates that on both datasets CJANet outperforms existing methods under various settings.

Comparison to the State-of-the-arts. CJANet outperforms GRU-D [Che *et al.*, 2018b], InterpNet [Shukla and Marlin, 2019], and DATA-GRU [Tan *et al.*, 2020a]. This is because these methods analyze health data in the time domain, which cannot tackle the discrepant dominant frequency issue. As a result, they are incapable of learning vital medical clues from varying sampling rates and fluctuation frequencies, which impair their results.

Effectiveness of Joint Attention. Variants with joint attention achieve better results than models without this mechanism. We can observe that GRU-m+j and GRU-p+j consistently outperform GRU-m and GRU-p, which indicate the effectiveness of the proposed joint attention. This is because the designed attention can identify importance scores of different features and frequencies, which provide important clues to increase the contributions of key elements. Furthermore, it provides clinical interpretability for the obtained risk prediction results, which is analyzed in the case study section.

Effectiveness of Cooperative Learning. We compare CJANet and Twin-GRU+jf (i.e., CJANet without cooperative learning). We can see that CJANet continuously outperforms



(a) Attention weight matrix learned for magnitude features

(b) Attention weight matrix learned for phase features

Figure 3: Visualization of attention weights. (a) and (b) provides attention weights learned for magnitude and phase signals, respectively.

Disease	Clinical outcome	Death time	Risk score predicted by CJANet
Stroke	Death	13.135 days	0.922

Table 3: Patient information in the case study.

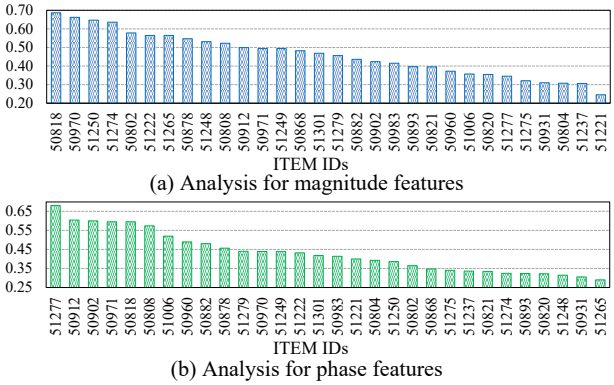
Twin-GRU+jf under various settings, which indicates the effectiveness of the cooperative learning module. These results indicate that this learning strategy can effectively enhance the information sharing capacity between magnitude and phase channels, enabling them to adaptively utilize global signals to optimize their parameters and promote results.

Effectiveness of Frequency-aware Fusion. We finally compare CJANet and Twin-GRU+jc (i.e., CJANet without frequency-aware fusion). We see that CJANet continually outperforms Twin-GRU+jc for various tasks, which proves the effectiveness of the frequency-aware fusion module. This is because this structure can adaptively learn the importance scores of different frequencies and automatically adjust their contributions. Therefore, influences of features at important frequencies are strengthened to promote final results.

4.4 Case Study

Fig. 3 provides a case study for predicting in-hospital mortality risk of a *Stroke* patient in MIMIC-III, whose information is given in Table 3. X-axis denotes frequencies and Y-axis represents different variables. Fig. 3(a) shows attention weights learned for magnitudes. We can observe that different variables contribute diverse weights and even for one same type of variable, the weights vary along frequency, which indicates that joint attention can effectively learn contribution scores of different variables and frequencies. By calculating mean of each variable at different frequencies, its average importance is obtained, as shown in Fig. 4(a). The variables assigned with the two largest weights are 50818 (*pCO2*) and 50970 (*Phosphate*), which have been proved to be important for evaluating the status of *Stroke* patients [Vats *et al.*, 2019]. This proves that CJANet can effectively identify key factors and provides vital references for practical medical applications.

Fig. 3(b) provides the attention weight matrix learned for phases, which indicates that contribution scores are adaptively adjusted for different features and frequencies. The average weight of each phase feature is provided in Fig. 4(b). The variables assigned with the two largest weights for the phase features are 51277 (*RDW*) and 50912 (*Creatinine*), which are vital for predicting severity of *Stroke* [Jia *et al.*, 2015]. These results prove that CJANet can also identify important phase variables and frequency ranges, which provide key clues for interpreting the achieved results and help design more effective treatments to improve outcomes. Furthermore, CJANet can jointly analyze magnitude and phase signals, successfully mining diverse yet key variables for accurate clinical predictions. These results prove that the proposed attention mechanism is able to analyze frequency signals from different aspects and mining diverse key medical knowledge, effectively improving final risk prediction results.



(a) Analysis for magnitude features

(b) Analysis for phase features

Figure 4: Average attention weights learned for each type of magnitude and phase features.

5 Conclusion

This paper proposes a novel CJANet to achieve accurate risk predictions by analyzing IMR-MTS in the frequency domain. Novel joint attention is designed to dynamically learn importance scores of different features at different frequencies, thus adaptively tackling discrepant data quality and dominant frequency bands. Furthermore, a new cooperative learning module is introduced to enhance information exchange between magnitudes and phases to utilize global signals to optimize the network. Extensive experimental results and case study indicate that CJANet significantly outperforms existing methods and provides highly interpretable prediction results.

Acknowledgements

This work was supported by the Health and Medical Research Fund Project under Grant 07180216.

References

[Breiman, 2001] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Che *et al.*, 2018a] Zhengping Che, Guangyu Li, et al. Hierarchical deep generative models for multi-rate multivariate time series. In *ICML*, pages 784–793, 2018.

[Che *et al.*, 2018b] Zhengping Che, Yan Liu, et al. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

[Choi *et al.*, 2016] Edward Choi, Jimeng Sun, et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, pages 3504–3512, 2016.

[Choi *et al.*, 2017] Edward Choi, Le Song, et al. Gram: graph-based attention model for healthcare representation learning. In *KDD*, pages 787–795, 2017.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[Fang *et al.*, 2020] Zhijie Fang, Weiqun Wang, Shixin Ren, et al. Learning regional attention convolutional neural network for motion intention recognition based on eeg data. In *IJCAI*, pages 1570–1576, 2020.

[Heo *et al.*, 2018] Jay Heo, Hae Beom Lee, Saehoon Kim, et al. Uncertainty-aware attention for reliable interpretation and prediction. In *NIPS*, pages 909–918, 2018.

[Hosmer Jr *et al.*, 2013] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[Issa *et al.*, 2020] Sali Issa, Qinmu Peng, and Xinge You. Emotion classification using eeg brain signals and the broad learning system. *IEEE TSMC: Systems*, 2020.

[Jia *et al.*, 2015] He Jia, Yan Zhang, et al. Association between red blood cell distribution width (rdw) and carotid artery atherosclerosis (cas) in patients with primary ischemic stroke. *Archives of Gerontology and Geriatrics*, 61(1):72–75, 2015.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[Li *et al.*, 2018] Shuai Li, Wanqing Li, Chris Cook, et al. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *CVPR*, pages 5457–5466, 2018.

[Lin and others, 2016] Sian-Jheng Lin et al. Novel polynomial basis with fast fourier transform and its application to reed–solomon erasure codes. *IEEE TIT*, 62(11):6284–6299, 2016.

[Neil *et al.*, 2016] Daniel Neil, Shih-Chii Liu, et al. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *NIPS*, pages 3882–3890, 2016.

[Pang *et al.*, 2020] Meng Pang, Qiquan Shi, et al. Iterative dynamic generic learning for face recognition from a contaminated single-sample per person. *IEEE TNNLS*, 2020.

[Pang *et al.*, 2021] Meng Pang, Binghui Wang, et al. Vdgan: A unified framework for joint prototype and representation learning from contaminated single sample per person. *IEEE TIFS*, 16:2246–2259, 2021.

[Parhi and Zhang, 2019] Keshab K Parhi and Zisheng Zhang. Discriminative ratio of spectral power and relative power features derived via frequency-domain model ratio with application to seizure prediction. *IEEE TBioCAS*, 13(4):645–657, 2019.

[Qiao *et al.*, 2018] Zhi Qiao, Cao Xiao, et al. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. In *IJCAI*, pages 3520–3526, 2018.

[Shukla and Marlin, 2019] Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. In *ICLR*, 2019.

[Tan *et al.*, 2018] Qingxing Tan, Andy Jinhua Ma, Huiqi Deng, Pong C Yuen, et al. A hybrid residual network and long short-term memory method for peptic ulcer bleeding mortality prediction. In *AMIA*, pages 998–1007, 2018.

[Tan *et al.*, 2020a] Qingxiong Tan, Mang Ye, Pong C Yuen, et al. Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *AAAI*, pages 930–937, 2020.

[Tan *et al.*, 2020b] Qingxiong Tan, Mang Ye, Pong C Yuen, et al. Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series. *IEEE TNNLS*, 2020.

[Tan *et al.*, 2021] Qingxiong Tan, Mang Ye, Andy Jinhua Ma, Pong C Yuen, et al. Importance-aware personalized learning for early risk prediction using static and dynamic health data. *JAMIA*, 28(4):713–726, 2021.

[Vats *et al.*, 2019] Kanchan Vats, Deepaneeta Sarmah, et al. Intra-arterial stem cell therapy diminishes inflammasome activation after ischemic stroke: a possible role of acid sensing ion channel 1a. *Journal of Molecular Neuroscience*, pages 1–8, 2019.

[Xu *et al.*, 2018] Yanbo Xu, Jimeng Sun, et al. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*, pages 2565–2573, 2018.

[Ye *et al.*, 2020] Mang Ye, Jianbing Shen, et al. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE TPAMI*, 2020.

[Zhang and Williamson, 2019] Michael Minyi Zhang and Sinead A Williamson. Embarrassingly parallel inference for gaussian processes. *JMLR*, 20:1–26, 2019.

[Zhang *et al.*, 2020] Gaobo Zhang, Zhen Mei, et al. A non-invasive blood glucose monitoring system based on smartphone ppg signal processing and machine learning. *IEEE TII*, 2020.

[Zhao *et al.*, 2020] Jingyu Zhao, Feiqing Huang, et al. Do rnn and lstm have long memory? In *ICML*, 2020.