Multi-View Integrative Attention-Based Deep Representation Learning for Irregular Clinical Time-Series Data

Yurim Lee, Eunji Jun, Jaehun Choi, and Heung-Il Suk, *Senior Member, IEEE*

***Abstract*—Electronic health record (EHR) data are sparse and irregular as they are recorded at irregular time intervals, and different clinical variables are measured at each observation point. In this work, to handle irregular multivariate time-series data, we consider the *human knowledge* of the aspects to be measured and time to measure them in different situations, known as multi-view features, which are indirectly represented in the data. We propose a scheme to realize multi-view features integration learning via a self-attention mechanism. Specifically, we devise a novel multi-integration attention module (MIAM) to extract complex information that is inherent in irregular time-series data. We explicitly learn the relationships among the observed values, missing indicators, and time interval between the consecutive observations in a simultaneous manner. In addition, we build an attention-based decoder as a missing value imputer that helps empower the representation learning of the interrelations among multi-view observations for the prediction task this decoder operates only in the training phase so that the final model is implemented in an imputation-free manner. We validated the effectiveness of our method over the public MIMIC-III and PhysioNet challenge 2012 datasets by comparing with and outperforming the state-of-the-art methods in three downstream tasks *i.e.*, prediction of the in-hospital mortality, prediction of the length of stay, and phenotyping. Moreover, we conduct a layer-wise relevance propagation (LRP) analysis based on case studies to highlight the explainability of the trained model.**

***Index Terms*—Electronic health records, bioinformatics, irregular time series modeling, deep learning, self-attention.**

I. Introduction

THE electronic health record (EHR) represents a collection of patients' health data such as coded diagnoses, vital signs, lab test records, procedures, and textual narratives. In the recent decade, significant progress has been made in developing deep learning models using multivariate EHR time series owing to the abundance of health big datasets. However, learning the appropriate EHR representations is challenging for many deep learning models as the EHR is irregularly sampled; in particular, the observations are measured at different time points determined by the type of measurement, patient's health status, and availability of clinical staff. Owing to the different number of observations and lack of temporal alignment across the data, machine learning models that assume a fixed-dimensional feature space cannot be applied.

The common deep learning approaches for handling the irregularly sampled time series, which are based on convolutional neural network (CNN), recurrent neural network (RNN), and recently attention-based methods, have demonstrated their superiority in healthcare target tasks. Specifically, CNN is usually applied to capture the local temporal characteristics of clinical data by using the predefined kernel sizes on a small chunk of EHR features to identify local motifs, *i.e.*, co-occurrences of diseases [1] and patient similarity [2]. However, although these approaches can model the local temporal dependencies in the predefined kernel size, the modeling of global dependencies is challenging.

RNN-based methods have been the defacto methodology to address clinical time series data, as RNNs can manage various lengths of sequential data. However, conventional RNN methods are designed to handle data with a constant time interval between consecutive time series, leading to suboptimal performance in the case of an irregular time interval. To address this challenge, the widely used approach is to convert irregularly sampled time series data into regularly sampled time series, *i.e.*, *temporal discretization* [3], [4] and feed this fixed-dimensional vector to the RNNs. Nonetheless, this process involves ad-hoc choices regarding the window size and aggregation function that address values falling in the same window. Similar to discretization methods, interpolation methods [4], [5] require specification of

Manuscript received 14 October 2021; revised 2 February 2022 and 6 April 2022; accepted 26 April 2022. Date of publication 5 May 2022; date of current version 9 August 2022. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korea government (MSIT) under Grants 2021-0-00982 for the Development of the multi-center distributed intelligence reinforcement federation technology for cooperating optimal treatments, and 2019-0-00079 for the Department of Artificial Intelligence (Korea University). *(Yurim Lee and Eunji Jun contributed equally to this work.) (Corresponding author: Heung-Il Suk.)*

Yurim Lee and Eunji Jun are with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: yurimalee@korea.ac.kr; ejjun92@gmail.com).

Jaehun Choi is with the Medical Information Research Section, Intelligent Convergence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea (e-mail: jhchoi@etri.re.kr).

Heung-Il Suk is with the Department of Artificial Intelligence and the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: hisuk@korea.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2022.3172549, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2022.3172549

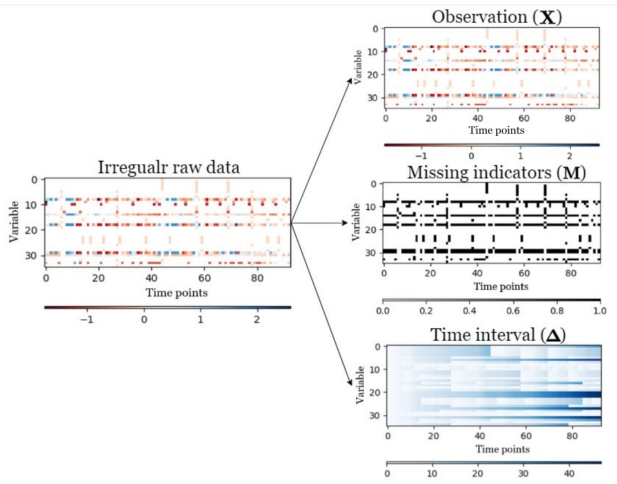

Fig. 1. Visualization of multi-view data. An observation ($\mathbf{X}$) represents the observed time series of all variables, missing indicators ($\mathbf{M}$) indicate whether the variables are observed (1) or missing (0), and the time interval ($\mathbf{\Delta}$) indicates the difference between successive observation time points.

discrete reference time points. In such approaches, the available observations in the input with the interpolants at these time points may inevitably introduce additional noise or information loss due to the assumption of a fixed time interval.

A more effective approach for handling irregular time series is to directly model the unequally spaced original data. Unlike the conventional RNN that relies on discrete time, ordinary differential equation(ODE)-based recurrence models [6] generalized the hidden state transitions in an RNN to continuous time dynamics. An alternative is to exploit the *source of missingness* such as missing indicators and time interval to model the informative missingness pattern. The existing works [3], [7]–[10] used either missing indicators or the time interval, and applied the heuristic decaying function such as a monotonically non-increasing function, without learning the representations for missingness.

Recently, attention-based methods [11]–[15] have been used to deal with irregular sampling instances. In particular, self-attention models [16] provide computational advantages over RNNs due to their fully parallelized sequence processing. Several works based on self-attention mechanism applied a simple modified self-attention scheme such as masked attention [14], or replaced the positional encoding with time encoding and concatenated encoding vectors and missing indicators [15].

To address the aforementioned limitations, we consider the *human knowledge* such as regarding the aspects to be measured and time to measure them in different situations, which are indirectly represented in the data in the shape of incompleteness and imperfections. These facets are known as multi-view features, *i.e.*, observations, missing indicators, and time intervals, as shown in Fig. 1. Specifically, we propose a novel method to jointly learn the deep representations of multi-view features from irregular multivariate time series data by using the self-attention mechanism in an imputation-free manner. The main contributions of this research can be summarized as follows:

- We devise a novel multi-view integration attention module (MIAM) to learn complex missing patterns by integrating the missing indicators and time interval, and further combine the observation and missing patterns in the representation space through a series of self-attention blocks.
- We build an attention-based decoder on top of the MIAM as a missing data imputer that helps empower the representation learning of the interrelations among multi-view observations for the prediction task; this imputer operates only in the training phase.
- We demonstrate that the proposed method can outperform state-of-the-art (SOTA) methods in downstream tasks, such as the prediction of the in-hospital mortality, prediction of the length of stay (LOS), and phenotyping on real-world EHR datasets, *i.e.*, the Medical Information Mart for the Intensive Care (MIMIC-III) and PhysioNet 2012 challenge datasets.
- We analyze the trained model by applying layer-wise relevance propagation (LRP) to demonstrate the explainability of the highlighted variables and time points related to the final prediction for an observation and its corresponding human knowledge information.

II. Related Work

1) Irregular Time Series Modeling: To accommodate irregularly sampled time series data, the widely-used approach is to discretize the time into consecutive and non-overlapping uniform intervals, *i.e.*, *temporal discretization* [3], [4], which enables the use of models that operate on fixed-dimensional vectors. Discretization reduces the problem from irregular time series data modeling to missing data imputation. Simple imputation approaches range from simple zero imputation and forward filling to more sophisticated deep learning approaches [17] including the use of a generative adversarial network (GAN) [18] and variational autoencoder (VAE) [19], [20], *etc.* However, the explicit imputation of the missing data during discretization is primarily based on heuristic or unsupervised methods that are not universally applicable, and cannot consider the uncertainty [21] in downstream clinical tasks.

Similar to discretization methods, interpolation methods require the specification of discrete reference time points. A multitask Gaussian process (MGP-RNN) model [4] conducted a probabilistic interpolation by transforming an irregular time series into a more uniform representation on evenly spaced reference time points, and feeding the latent function values into RNNs. Although this approach provided uncertainty, deficiencies exist in terms of the predefined time interval between the reference time points and limited expressiveness of the model in terms of the sum of separable kernel functions. In contrast, the interpolation-prediction network (IPNet) [5] learns an optimal time interval to perform deterministic interpolation; first by interpolating the irregular time series for each variable, and later merging all-time series across every variable. However, IPNet may unavoidably introduce additional noise or information loss because a fixed time interval is assumed.

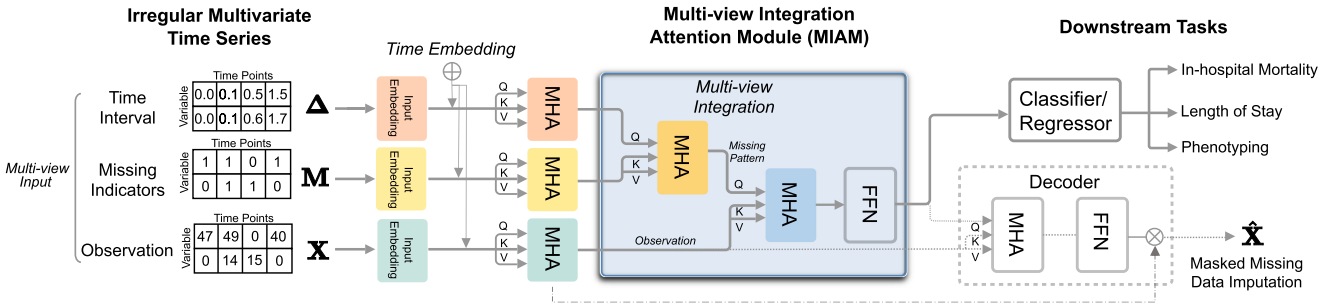

Fig. 2. Overall framework of the proposed multi-view integration learning method for irregularly-sampled clinical time series. The solid lines represent the classification process, and dotted lines represent the auxiliary missing data imputation process. For the input embedding and multihead self-attention (MHA), each observation and missing indicators have individual learnable weights, and two integration MHAs have separate weights.

Accordingly, the recent interest in irregularity-based methods is learning representations directly from multivariate sparse and irregularly sampled time series as the input without the need for imputation [6], [15], [22]. Specifically, ODE-LSTM [6] manages continuous time observations within the LSTM network, enabling cells to handle nonuniform time intervals and eliminate the need to aggregate observations into equally spaced intervals.

2) Missing Pattern Modeling: To model informative missingness, Che *et al.* [7] added a temporal decay derived from a time interval to the input variables and hidden states and directly incorporated both the observation and missing indicators in the GRU architecture. Furthermore, Lipton *et al.* [3] used hand-engineered features derived from the response indicator, such as the mean and standard deviation of the missing indicators, for each time series. Baytas *et al.* [8] proposed a time-aware LSTM (T-LSTM) to decompose the cell memory into short and long-term memories and decay the short-term memory through the weights transformed from time intervals, while retaining the long-term memories. Specifically, a monotonically nonincreasing function was heuristically selected as a decaying function. Similarly, ATTAIN [9] decayed the short-term memory using both time intervals and weights generated from the attention mechanism by considering several previous events, instead of only previous event. DATA-GRU [10] introduced a time-aware mechanism to a GRU for handling irregular time intervals, and further devised a dual attention mechanism to address missing values in both data-quality and medical-knowledge views.

3) Attention Mechanism in Irregular Time Series Modeling: Several recent models leveraged attention mechanisms as the fundamental approach to deal with irregular sampling. For example, RETAIN [11] learned an interpretable representation of irregularly sampled time series through two-level RNNs, and generated visit-level and variable-level attentions. To learn the robust representations of EHR data, Choi *et al.* [12] introduced a graph-based attention method. Similarly, Ma *et al.* [13] proposed a knowledge-based attention mechanism to learn the embeddings for nodes in the knowledge graph. Based on the self-attention [16], Song *et al.* [14] proposed a SAnD model that adopted a masked self-attention to specify how far the attention model looks into the past and a dense interpolation to capture the temporal dependencies. In addition, Horn *et al.* [15] replaced the positional encoding with time encoding and concatenated encoding vectors and missing indicators.

III. Method

In this section, we present the proposed method for multi-view feature integration learning of irregular multivariate EHR time series for downstream tasks, *i.e.*, in-hospital mortality prediction, LOS prediction, and phenotyping. First, we introduce the notations for multivariate time series data, and subsequently describe the proposed method that consists of (i) input and time embedding, (ii) multi-view integration learning, (iii) downstream prediction tasks, and (iv) auxiliary imputation for masked missing data. The overall architecture is shown in Fig. 2.

A. Data Representation

For each subject n, given a set of D-dimensional multivariate time series in $\mathbf{t}^{(n)} = [t_1, \ldots, t_j, \ldots, t_{T_n}]$ time points of length T_n, we denote an observation time series as $\mathbf{X}^{(n)} = [\mathbf{x}_{t_1}^{(n)}, \ldots, \mathbf{x}_{t_j}^{(n)}, \ldots, \mathbf{x}_{t_{T_n}}^{(n)}]^\top \in \mathbb{R}^{T_n \times D}$, where $\mathbf{x}_{t_j}^{(n)} \in \mathbb{R}^D$ represents the t_j-th observation of all variables, and $x_{t_j,d}^{(n)}$ is the element of the d-th variable in $\mathbf{x}_{t_j}^{(n)}$. In this setting, because the time series $\mathbf{X}^{(n)}$ includes missing values, we introduce a missing indicator across the time series, $\mathbf{M}^{(n)} = [\mathbf{m}_{t_1}^{(n)}, \ldots, \mathbf{m}_{t_j}^{(n)}, \ldots, \mathbf{m}_{t_{T_n}}^{(n)}]^\top \in \mathbb{R}^{T_n \times D}$, which has the same size as $\mathbf{X}^{(n)}$ to mark the variables that are observed or missing. Specifically, $m_{t_j,d}^{(n)} = 1$ is observed if $x_{t_j,d}^{(n)}$ is observed, otherwise, $m_{t_j,d}^{(n)} = 0$. If the observation is missing, the input is set as zero for the corresponding elements in $\mathbf{x}_{t_j}^{(x)}$. For each variable d, we present a *time interval* $\boldsymbol{\Delta}^{(n)} = [\boldsymbol{\delta}_{t_1}^{(n)}, \ldots, \boldsymbol{\delta}_{t_j}^{(n)}, \ldots, \boldsymbol{\delta}_{t_{T_n}}^{(n)}]^\top \in \mathbb{R}^{T_n \times D}$, where $\delta_{t_j,d}^{(n)} \in \mathbb{R}$ is defined as:

$$\boldsymbol{\delta}_{t_j,d}^{(n)} = \begin{cases} t_j - t_{j-1} + \boldsymbol{\delta}_{t_{j-1},d}^{(n)}, & t_j > 1, m_{t_{j-1}^{(n)},d} = 0 \\ t_j - t_{j-1}, & t_j > 1, m_{t_{j-1}^{(n)},d} = 1 \\ 0, & t_j = 1. \end{cases} \quad (1)$$

In this study, given a clinical time series dataset $\mathcal{D} = \{(\mathbf{X}^{(n)}, \mathbf{M}^{(n)}, \mathbf{\Delta}^{(n)})\}_{n=1}^{N}$ for N subjects, we construct a mapping function to define the prediction labels $(y_1, \ldots, y_N)$.

For uncluttered, we will use functional notation that represents information regarding a particular patient, omitting the superscript (n) for the n-th subject.

B. Multi-View Integration Learning

The key feature of missing data is that the missingness may convey meaningful information, and ignoring this dependence may lead to incorrect predictions. The existing works [3], [7]–[10] leveraged these sources of missingness, *i.e.*, missing indicators and time interval, and applied the heuristic decaying function to use these aspects without learning their representations. However, the inappropriate modeling of missingness may lead to the unreliable assessment of the feature importance and the generation of a model that is not robust to measurement changes.

Considering these aspects, in this work, we learn a deep representation of the irregular time series data by effectively leveraging both missing indicators and time intervals. We consider these sources of missingness as human knowledge in terms of what to measure and when to measure it in different situations; these aspects are indirectly represented in the data. In this context, we regard the representations of the missing indicators and time interval as multi-view features of irregularly sampled observation. Specifically, we propose a multi-view feature integration learning for modeling the interrelations among multi-view observations. To this end, a self-attention mechanism is used, in which the inner product of representations often reflects relationships such as a similarity.

1) Input and Time Embedding: Given the D measurements at every time step, the first step is to learn the respective input embeddings for the observation, missing indicators, and time interval. Compared to an existing approach [16] that considered only the sequence order instead of the temporal patterns in positional encoding, we employ a time embedding (TE) as a variant of positional encoding that takes continuous time values as input, and convert them into an encoding vector representation. This approach deals with irregularly sampled time series by considering the exact time points and their time interval. For time embedding, the sine and cosine functions proposed in [16] are modified as follows:

$$\mathrm{TE}_{(t,d)} = \begin{cases} \sin\left(\frac{t}{L_{\max}^{2\,d/d_{\text{model}}}}\right) & \text{if } d \text{ is even} \\ \cos\left(\frac{t}{L_{\max}^{2\,d/d_{\text{model}}}}\right) & \text{if } d \text{ is odd} \end{cases} \tag{2}$$

where t, d, d_{model}, and $L_{\max}$ denote the exact time point, an index of the variable, the model dimension, and the maximum time length of data, respectively. The time embeddings are added to the learned input embeddings:

$$\mathbf{X}^* = \mathbf{W}_{\mathbf{X}}\mathbf{X} + \mathrm{TE}(t,d) \tag{3}$$

$$\mathbf{M}^* = \mathbf{W}_{\mathbf{M}}\mathbf{M} + \mathrm{TE}(t,d) \tag{4}$$

$$\mathbf{\Delta}^* = \mathbf{W}_{\mathbf{\Delta}}\mathbf{\Delta} + \mathrm{TE}(t,d) \tag{5}$$

where $\mathbf{W}_{\mathbf{X}}$, $\mathbf{W}_{\mathbf{M}}$, and $\mathbf{W}_{\mathbf{\Delta}}$ are corresponding embedding weight matrices for the respective multi-view inputs.

2) Self-Attention: The basic building block for this approach is based on multi-head self-attention (MHA) [16], in which a scaled dot-product attention is calculated over a set of queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$). Based on the self-attention block, we learn the attention representations of multi-view irregular time series including the observation $\mathbf{X}$, masking indicators $\mathbf{M}$, and time interval $\mathbf{\Delta}$. Specifically, each input set ($\mathbf{X}^*$, $\mathbf{M}^*$, $\mathbf{\Delta}^*$) learns its own representation ($\mathbf{H}^{\mathbf{X}}$, $\mathbf{H}^{\mathbf{M}}$, $\mathbf{H}^{\mathbf{\Delta}}$) through self-attention block, in which each data point is linearly combined with its own weight matrix and fed to the corresponding $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$:

$$\mathbf{H}^{\mathbf{X}} = \sigma\left(\frac{(\mathbf{W}^{\mathbf{Q}}_{\mathbf{X}^*}\mathbf{X}^*)(\mathbf{W}^{\mathbf{K}}_{\mathbf{X}^*}\mathbf{X}^*)^\top}{\sqrt{d_k}}\right)(\mathbf{W}^{\mathbf{V}}_{\mathbf{X}^*}\mathbf{X}^*) \tag{6}$$

$$\mathbf{H}^{\mathbf{M}} = \sigma\left(\frac{(\mathbf{W}^{\mathbf{Q}}_{\mathbf{M}^*}\mathbf{M}^*)(\mathbf{W}^{\mathbf{K}}_{\mathbf{M}^*}\mathbf{M}^*)^\top}{\sqrt{d_k}}\right)(\mathbf{W}^{\mathbf{V}}_{\mathbf{M}^*}\mathbf{M}^*) \tag{7}$$

$$\mathbf{H}^{\mathbf{\Delta}} = \sigma\left(\frac{(\mathbf{W}^{\mathbf{Q}}_{\mathbf{\Delta}^*}\mathbf{\Delta}^*)(\mathbf{W}^{\mathbf{K}}_{\mathbf{\Delta}^*}\mathbf{\Delta}^*)^\top}{\sqrt{d_k}}\right)(\mathbf{W}^{\mathbf{V}}_{\mathbf{\Delta}^*}\mathbf{\Delta}^*) \tag{8}$$

where $\mathbf{W}$ are a set of the learnable weight matrices, σ is SoftMax activation function, and d_k is the dimension of the key vector.

3) Multi-View Integration Attention: In this work, we devise a novel multi-view integration attention module (MIAM) that consists of two submodules: (i) an integration module that relies mostly on the self-attention mechanism, and (ii) a position-wise fully connected feed-forward network (FFN) module. The integration module aims to learn a complex missing pattern by integrating the missing indicators and time interval, and further combines the observation and learned missing pattern in the representation space. While most works in the literature [3], [7]–[10] exploited either the missing indicators or time interval and applied the heuristic decaying function to enable their use, we effectively learn the informative missing pattern by using both the missing indicators and time interval in the representation space. We argue that learning the underlying representation from the missingness itself eliminates the need to impute values, and does not require the specification of any heuristic function.

The integration module involves two integration steps, *i.e.*, missingness integration and observation-missingness integration. In the first step, we incorporate the representation of the missing indicators with that of the time interval by self-attention block, thereby obtaining the representation of missing pattern ($\mathbf{H}^{\text{miss}}$).

$$\mathbf{H}^{\text{miss}} = \sigma\left(\frac{(\mathbf{W}^{\mathbf{Q}}_{\mathbf{\Delta}}\mathbf{H}^{\mathbf{\Delta}})(\mathbf{W}^{\mathbf{K}}_{\mathbf{M}}\mathbf{H}^{\mathbf{M}})^\top}{\sqrt{d_k}}\right)(\mathbf{W}^{\mathbf{V}}_{\mathbf{M}}\mathbf{H}^{\mathbf{M}}) \tag{9}$$

Similarly, in the second step, the representation of the observation is combined with that of the missing pattern through another self-attention block to incorporate the information of the missing pattern to that of variables.

$$\mathbf{H}^{\text{obs-miss}} = \sigma\left(\frac{(\mathbf{W}^{\mathbf{Q}}_{\text{miss}}\mathbf{H}^{\text{miss}})(\mathbf{W}^{\mathbf{K}}_{\mathbf{X}}\mathbf{H}^{\mathbf{X}})^\top}{\sqrt{d_k}}\right)(\mathbf{W}^{\mathbf{V}}_{\mathbf{X}}\mathbf{H}^{\mathbf{X}}) \tag{10}$$

This final attention output is the jointly learned deep representations that model the relation between the irregular observation data and missing pattern. Subsequently, the FFN module is applied to each time point identically to model the dependency among variables. The final representation $\tilde{\mathbf{H}}$ is leveraged for the downstream prediction and auxiliary imputation tasks.

C. Downstream Prediction Tasks

After obtaining the final representation from the MIAM, we construct a classifier and regressor to accomplish each of the three downstream tasks, *i.e.*, (i) prediction of in-hospital mortality, (ii) prediction of LOS, and (iii) phenotyping. In-hospital mortality pertains to a binary classification that indicates whether a patient dies during the period of hospital admission or lives to be discharged. The target label for the patients is $\mathbf{y} = (y_1, \ldots, y_N | y_n \in \{0, 1\})$. LOS prediction is a regression task in which the number of days between the patient's admission to the ICU and end of hospitalization is predicted. Phenotyping represents a multi-label classification problem that shows patients' diagnosis as a patient is diagnosed with multiple conditions.

To predict the probability of downstream tasks, given the final representation $\tilde{\mathbf{H}}$, we conduct an average pooling over the timestamps, which results in a final pooled representation $\hat{\mathbf{h}}$, followed by a multi-layer perceptron (MLP). The final layer depends on the specific task, and we a sigmoid layer is used for for classification and multilabel classification.

Furthermore, to address the poor classification performance problem in highly imbalanced data found in healthcare datasets, we employ focal loss [21], [23] as the objective function to calculate the classification loss between the target label y and predicted label $\hat{y}$ of each patient. As γ and β of the focal loss smoothly adjust the weighting rate of easy or hard examples. Hence, for the three downstream tasks, the corresponding loss functions $\mathcal{L}_{\text{pred}}$ are defined as follows:

- Binary classification:

$$\frac{1}{N}\sum_{n=1}^{N} -\beta(1-\hat{y}^{(n)})^{\gamma}\log(\hat{y}^{(n)}) \tag{11}$$

- Regression:

$$\sqrt{\frac{1}{N}\sum\nolimits_{n=1}^{N}(\hat{y}^{(n)}-y^{(n)})^2} \tag{12}$$

- Multi-label classification:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K} -\beta\left(1-\hat{y}_k^{(n)}\right)^{\gamma}\log\left(\hat{y}_k^{(n)}\right)\right) \tag{13}$$

where N is the total number of patients, K is the total number of labels in the dataset, γ is a focusing parameter for a minority class, and β is a weighting factor to balance the importance between classes.

D. Auxiliary Missing Data Imputation

On top of the MIAM module, we build an attention-based decoder as a missing data imputer that aims to enhance the representational power of the inter-relations among multi-view observations for the prediction task. To investigate the masked imputation loss, we randomly mask 10% of the non-missing values and predict them. From a self-supervised learning perspective, this task can be regarded as similar to a masked language modeling task accomplished using BERT [24], which randomly masks several tokens in a text sequence and independently recovers the masked tokens to learn the language representations. By taking a similar approach, we learn the interrelations between the corrupted values and context, which further contributes to the learning of the interrelations among multi-view observations. Notably, the proposed method is basically an imputation-free method as the imputation is implemented only in the training phase and not the test phase. Therefore, the model complexity is reduced, while avoiding the existing imputation-related problems that lead to strong biases due to improper imputation estimates.

In the attention-based decoder, we further apply an attention block between the final representation of the MIAM ($\tilde{\mathbf{H}}$) as a query and observation representation ($\mathbf{H}^{\mathbf{X}}$) as the key and value, followed by the FFN. The output layer maps the output of FFN to the target time sequence by utilizing the learned embedding matrix ($\mathbf{W}_{\mathbf{X}}$), and yields the imputed data ($\hat{\mathbf{X}}$), as shown in (14).

$$\hat{\mathbf{X}} = \mathbf{W}_{\mathbf{X}}^{\top} \times \text{FFN}\left(\sigma\left(\frac{\left(\mathbf{W}_{\tilde{\mathbf{H}}}^{\mathbf{Q}}\tilde{\mathbf{H}}\right)\left(\mathbf{W}_{\hat{\mathbf{X}}}^{\mathbf{K}}\mathbf{H}^{\mathbf{X}}\right)^{\top}}{\sqrt{d_k}}\right)\left(\mathbf{W}_{\hat{\mathbf{X}}}^{\mathbf{V}}\mathbf{H}^{\mathbf{X}}\right)\right) \tag{14}$$

Given another masking vector $\mathbf{M}_{\text{imp}}$ introduced for the purpose of marking the masked values, the imputation loss $\mathcal{L}_{\text{imp}}$ is calculated by *masked mean squared error (MSE)* between the original sample $\mathbf{X}$ as the ground truth and imputed sample $\hat{\mathbf{X}}$

Algorithm 1: Algorithm of Our Proposed Model for in-hospital Mortality Prediction.

input : irregular clinical time series data $\{\mathbf{X}, \mathbf{y}, \mathbf{M}, \mathbf{\Delta}, \mathbf{t}\}$
output: outcome prediction $\hat{\mathbf{y}}$
while *not converge* **do**
  # Input & time embedding:
  $\mathbf{X}^*, \mathbf{M}^*, \mathbf{\Delta}^* \leftarrow$ Eq. (3), (4), (5)
  # MIAM:
  $\mathbf{H}^{\mathbf{X}}, \mathbf{H}^{\mathbf{M}}, \mathbf{H}^{\mathbf{\Delta}} \leftarrow$ Eq. (6), (7), (8)
  $\mathbf{H}^{\text{miss}} \leftarrow$ Eq. (9)
  $\mathbf{H}^{\text{obs-miss}} \leftarrow$ Eq. (10)
  $\tilde{\mathbf{H}} \leftarrow \text{FFN}(\mathbf{H}^{\text{obs-miss}})$
  # Prediction:
  $\hat{\mathbf{h}} = \text{AvgPool}(\tilde{\mathbf{H}})$
  $\hat{\mathbf{y}} \leftarrow \sigma(\text{LeakyReLU}(\mathbf{W}_1\hat{\mathbf{h}} + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2)$
  $\mathcal{L}_{\text{pred}} \leftarrow \text{FocalLoss}(\hat{\mathbf{y}}, \mathbf{y})$
  # Auxiliary Imputer:
  $\hat{\mathbf{X}} \leftarrow$ Eq. (14)
  $\mathcal{L}_{\text{imp}} \leftarrow$ Eq. (15)
  $\mathcal{L} \leftarrow \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{imp}}\mathcal{L}_{\text{imp}}$
  Optimize with RAdam
end

for the marked values only by $\mathbf{M}_{\text{imp}}$ as:

$$\mathcal{L}_{\text{imp}} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{X}^{(n)} \odot \mathbf{M}_{\text{imp}}^{(n)} - \hat{\mathbf{X}}^{(n)} \odot \mathbf{M}_{\text{imp}}^{(n)}\right)^2. \quad (15)$$

By introducing the imputation loss in the objective function, the imputer provides auxiliary information to achieve the optimal prediction result.

E. Composite Loss

The composite loss is defined by accumulating the prediction and imputation losses as $\mathcal{L} = \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{imp}}\mathcal{L}_{\text{imp}}$ where λ_{prd} and λ_{imp} are hyperparameters that control the ratio between two losses. We optimize all the parameters of our model in an end-to-end manner via the composite loss $\mathcal{L}$. The overall training process the proposed method for the in-hospital mortality task is summarized as Algorithm 1. The auxiliary imputer operates only in the training phase.

IV. Experimental Settings

In this section, we evaluated our proposed method for the downstream tasks illustrated in Section III-C. Each of the tasks is learned separately in a single-task manner.

1) Dataset and Preprocessing: We conducted experiments on two datasets: MIMIC-III[1] and PhysioNet 2012 challenge[2] datasets. MIMIC-III dataset consists of medical records of 13,998 patients collected at Beth Israel Deaconess Medical Center. We used 99 time series measurements for each patient. The number of irregular time points ranged from 1 to 247 ($m \pm \sigma = 49.29 \pm 35.90$). The selected time series was scarcely observed leading to a missing rate of approximately 90%. In terms of the in-hospital mortality label, the ratio of 1,181 positive (dead in hospital) to 12,817 negative (alive in-hospital) was approximately 1:10.8, and the LOS ranged from 0.0004 to 86.85 days ($m \pm \sigma = 3.81 \pm 4.86$). The phenotyping label included 25 disease conditions with 12, 8, 5 conditions being critical (such as respiratory/renal failure), chronic (such as diabetes or atherosclerosis), mixed conditions, respectively.

The PhysioNet 2012 challenge dataset consists of 35 different time-series measurements and five general descriptors for 4,000 critical care patients with at least 48 hours of hospital stay. We used 35 time-series variables without the general descriptor of 3,997 patients with at least one observation time. The number of irregular time points ranged from 1 to 202 ($m \pm \sigma = 73.87 \pm 23.06$). The missing rate was approximately 80.5%, the mortality labels were imbalanced at a ratio of approximately 1:6 between 554 positive and 3,443 negative cases, and LOS label varied from 0 to 105 ($m \pm \sigma = 13.45 \pm 12.24$).

In terms of data preprocessing, because each variable had a different range, all the inputs were first winsorized to remove outliers and then z-normalized using the global mean and standard deviation from the entire training set to achieve a zero mean and unit variance in a variable-wise manner.

[1]Available at https://mimic.physionet.org/.

[2]Available at https://physionet.org/content/challenge-2012/.

2) Performance Evaluation: We reported the average results from 5-fold cross validation, in comparison with other SOTA methods in the literature. For the in-hospital mortality prediction, *i.e.*, binary classification, we used the area under the ROC curve (AUC) and area under the precision–recall curve (AUPRC) as the evaluation metrics. To forecast LOS, *i.e.*, the regression task, we used the root mean squared error (RMSE) and mean absolute error (MAE) as the evaluation metrics. For phenotyping, *i.e.*, multi-label classification, we used the mean AUC (mAUC) and mean AUPRC (mAUPRC).

In addition, we conducted exhaustive ablation studies to investigate the influence of different experimental design options in terms of the multi-view integration module and auxiliary imputer in the proposed method for each task. We validated the effectiveness of multi-view integration by considering five scenarios, *i.e.*, single view ($\mathbf{X}$), double views ($\mathbf{M} + \mathbf{\Delta}$, $\mathbf{X} + \mathbf{M}$, $\mathbf{X} + \mathbf{\Delta}$), and triple views ($\mathbf{X} + \mathbf{M} + \mathbf{\Delta}$).

3) Experimental Settings: We trained our models using the rectified Adam (RAdam) optimizer [25] with an initial learning rate of 0.005 and multiplicative decay of 0.2 for every 10 epochs for 60 epochs with early stopping of 20 patience by using mini-batches of 64 samples. We chose the final optimal model based on the performance on the validation set.

For the MIMIC-III dataset, the model achieved its best performance for binary and multi-label classification at self-attention block with eight heads with $d_k = d_v = 128$, and $d_{FFN} = 128$, where d_k, d_v, d_{FFN} are the dimension of the key, value, and FFN, respectively, in all the self-attention blocks of the MIAM and imputer. For regression, the highest performance was achieved using a self-attention block with eight heads of $d_k = d_v = 64$, and $d_{FFN} = 128$. For the PhysioNet 2012 challenge dataset, we observed the highest performance at with a self-attention block with eight heads of $d_k = d_v = 64$, and $d_{\text{FFN}} = 128$ for binary and multi-label classification. For regression, the model exhibited the highest performance with at self-attention block with four heads $d_k = d_v = 64$, and $d_{FFN} = 64$. The β and γ in focal loss were chosen 7 and 0.15, respectively. The composite loss parameters λ_{imp} and λ_{pred} for the classification task were 10 and 1, respectively. In the regression task, λ_{imp} and λ_{prd} were both 1. For the activation function, we used LeakyReLU and Tanh for binary and multi-label classification, respectively.

We validated the efficacy of the proposed method by comparing it with SOTA methods categorized as recurrence-based, interpolation-based, and attention-based methods as shown in Table I and Table II.

V. Experimental Results and Analysis

Experimental Results

Table I and Table II compare the results of the proposed method with those of the competing methods for mortality prediction on the PhysioNet 2012 challenge and MIMIC-III datasets. The results of the ablation studies for five scenarios, *i.e.*, single view ($\mathbf{X}$), double views ($\mathbf{M} + \mathbf{\Delta}$, $\mathbf{X} + \mathbf{M}$, $\mathbf{X} + \mathbf{\Delta}$), and triple views ($\mathbf{X} + \mathbf{M} + \mathbf{\Delta}$), are presented in Table III and Table IV. The effect of the attention-based imputer was explored in Table V and Table VI. Lastly, we compared the proposed

TABLE I
RESULTS OF THE PROPOSED METHOD IN PHYSIONET 2012 CHALLENGE DATASET AND THE COMPETING METHODS FOR DOWNSTREAM TASKS ($m \pm \sigma$ FROM 5-FOLD CROSS-VALIDATION)

Category	Method	In-hospital Mortality		Length of Stay	
		AUC	AUPRC	RMSE	MAE
Recurrence-based	T-LSTM [8]	0.8012±0.0205	0.3932±0.0463	19.9576±0.8807	9.9276±0.4016
	GRU-D [7]	0.7990±0.0223	0.4215±0.0261	14.3913±0.8462	8.0931±0.4811
	ATTAIN [9]	0.8093±0.0318	0.4027±0.0394	13.2144±0.8079	7.6912±0.1574
	DATA-GRU [10]	0.7982±0.0160	0.3977±0.0509	15.9280±0.5539	8.9711±0.5619
	ODE-LSTM [6]	0.8065±0.0395	0.4190± 0.0425	13.1217±0.9226	7.6873±0.6743
Interpolation-based	MGP-RNN [4]	0.7970±0.0244	0.3754±0.0479	16.2858±0.6335	8.6608±0.3254
	IPNet [5]	0.7518±0.0295	0.3063±0.0177	17.1042±0.8020	8.8702±0.5245
Attention-based	SAnD [14]	0.7911±0.0373	0.4327±0.0543	13.3599±0.6243	8.0471±0.2825
	Proposed	**0.8201±0.0218**	**0.4698±0.0563**	**11.5743±0.8937**	**7.4762±0.1897**

TABLE II
RESULTS OF THE PROPOSED METHOD IN MIMIC-III DATASET AND THE COMPETING METHODS FOR DOWNSTREAM TASKS ($m \pm \sigma$ FROM 5-FOLD CROSS-VALIDATION)

Category	Method	In-hospital Mortality		Length of Stay		Phenotyping	
		AUC	AUPRC	RMSE	MAE	mAUC	mAUPRC
Recurrence-based	T-LSTM [8]	0.8216±0.0205	0.3475±0.0430	9.0428±0.0944	7.9853±0.0617	0.8094±0.0242	0.5196±0.0481
	GRU-D [7]	0.8371±0.0163	0.3662±0.0120	4.8248±0.0891	3.7037±0.0628	0.8232±0.0080	0.5492±0.0094
	ATTAIN [9]	0.8302±0.0312	0.3572±0.0304	5.2775±0.0907	3.1673±0.0769	0.8201±0.0364	0.5414±0.0338
	DATA-GRU [10]	0.8378±0.0117	0.3584±0.0153	6.7927±0.0997	5.0490±0.0892	0.8194±0.0076	0.5428±0.0097
	ODE-LSTM [6]	0.8423±0.0077	0.3513±0.0147	5.7900±0.0857	4.2862±0.0437	0.8189±0.0025	0.5254±0.0046
Interpolation-based	MGP-RNN [4]	0.8115±0.0194	0.3354±0.0377	5.9056±0.0534	4.2927±0.0729	0.8053±0.0235	0.5123±0.0438
	IPNet [5]	0.8226±0.0261	0.3592±0.0525	8.9984±0.0892	8.0080±0.0237	0.8037±0.0035	0.5060±0.0103
Attention-based	SAnD [14]	0.8256±0.0159	**0.3712±0.0220**	5.3559±0.0654	3.6915±0.0713	0.8154±0.0037	0.5202±0.0037
	Proposed	**0.8534±0.0071**	0.3565±0.0133	**4.4336±0.0859**	**2.0963±0.0859**	**0.8359±0.0034**	**0.5641±0.0085**

TABLE III
RESULTS OF ABLATION STUDIES FOR THE PHYSIONET 2012 CHALLENGE DATASET IN TERMS OF THE MULTI-VIEW INTEGRATION ($m \pm \sigma$ FROM 5-FOLD CROSS-VALIDATION)

Task	Measure	Method				
		Single ($\mathbf{X}$)	Double ($\mathbf{M} + \mathbf{\Delta}$)	Double ($\mathbf{X} + \mathbf{M}$)	Double ($\mathbf{X} + \mathbf{\Delta}$)	Triple ($\mathbf{X} + \mathbf{M} + \mathbf{\Delta}$)
Mortality	AUC	0.7816±0.0381	0.8110±0.0153	0.8131±0.0292	0.8142±0.0175	**0.8201±0.0218**
	AUPRC	0.4195±0.0345	0.4429±0.0211	0.4475±0.0452	0.4501±0.0225	**0.4698±0.0563**
LOS	RMSE	14.3701±0.7322	13.8129±0.7838	13.6428±0.8722	13.7811±0.7675	**11.5743±0.8937**
	MAE	8.5277±0.2301	8.2870±0.1311	7.5185±0.2397	8.2101±0.1247	**7.4762±0.1897**

The Bold Faced Text Pertains to the Proposed Approach

TABLE IV
RESULTS OF ABLATION STUDIES FOR THE MIMIC-III DATASET IN TERMS OF THE MULTI-VIEW INTEGRATION ($m \pm \sigma$ FROM 5-FOLD CROSS-VALIDATION)

Task	Measure	Method				
		Single ($\mathbf{X}$)	Double ($\mathbf{M} + \mathbf{\Delta}$)	Double ($\mathbf{X} + \mathbf{M}$)	Double ($\mathbf{X} + \mathbf{\Delta}$)	Triple ($\mathbf{X} + \mathbf{M} + \mathbf{\Delta}$)
Mortality	AUC	0.7943±0.0048	0.8261±0.0064	0.8273±0.0058	0.8458±0.0051	**0.8534±0.0071**
	AUPRC	0.3044±0.0091	0.3297±0.0139	0.3385±0.0170	0.3557±0.0155	**0.3565±0.0133**
LOS	RMSE	7.8655±0.0834	5.2822±0.0781	5.2776±0.0868	5.2736±0.0793	**4.4336±0.0859**
	MAE	5.2122±0.0804	3.1713±0.0726	2.9738±0.0902	3.1671±0.0764	**2.0963±0.0859**
Phenotyping	mAUC	0.7785±0.0039	0.8155±0.0033	0.81856±0.0029	0.8201±0.0038	**0.8359±0.0034**
	mAUPRC	0.4807±0.0086	0.4784±0.0079	0.49442±0.0071	0.5239±0.0074	**0.5641±0.0085**

The Bold Faced Text Pertains to the Proposed Approach

approach with the case of using the imputed data by the BRITS imputer [17], which is the SOTA imputation method, and the auxiliary imputer.

1) In-Hospital Mortality Prediction: On both the datasets, our model achieved the best classification performance with AUC of 0.8201 and 0.8534. Overall the recurrence-based methods outperformed the interpolation methods. Among the recurrence-based methods, ODE-LSTM [6] which takes irregularly sampled time series as input and simultaneously handles time intervals demonstrated a competitive performance. This finding suggests the need to directly model irregularly sampled time series and the more sophisticated use of missing indicators and time interval, as implemented in the proposed model.

2) LOS Prediction: The proposed model again demonstrated the highest performance achieving the lowest RMSE and MAE for both datasets. Most recurrence-based and attention-based methods showed relatively better performance than the imputation-based methods. ATTAIN [9] exhibited the highest

TABLE V
ABLATION STUDIES RELATED TO THE MISSING PATTERN GENERATION FOR THE PHYSIONET 2012 CHALLENGE DATASET

Task	Measure	Method	
		w/o Imputer	w/ Imputer
Mortality	AUC	0.8137±0.0232	**0.8201±0.0218**
	AUPRC	0.4515±0.0416	**0.4698±0.0563**
LOS	RMSE	12.0026±0.9475	**11.5743±0.8937**
	MAE	7.9691±0.2304	**7.4762±0.1897**

TABLE VI
ABLATION STUDIES RELATED TO THE ATTENTION-BASED IMPUTER FOR THE MIMIC-III DATASET

Task	Measure	Method	
		w/o Imputer	w/ Imputer
Mortality	AUC	0.8373±0.0077	**0.8534±0.0071**
	AUPRC	0.3487±0.0147	**0.3565±0.0133**
LOS	RMSE	6.5856±0.0875	**4.4336±0.0859**
	MAE	4.6810±0.0845	**2.0963 ±0.0859**
Phenotyping	mAUC	0.82916±0.0039	**0.8359±0.0034**
	mAUPRC	0.54849±0.0064	**0.5641±0.0085**

and second highest performance over the MIMIC-III dataset and PhysioNet 2012 challenge datasets, respectively. This finding highlights the need for adjusting the time interval instead of using this information directly. Notably, T-LSTM [8] showed the highest RMSE and MAE in both PhysioNet 2012 and MIMIC-III datasets, suggesting the necessity of missing pattern information, as the T-LSTM only considers the time intervals, unlike the other methods.

3) Phenotyping: The phenotyping result is available only for the MIMIC-III dataset as the PhysioNet 2012 challenge dataset does not provide phenotyping labels. The attention-based methods outperformed the recurrence-based and imputation-based methods. In particular, our proposed method showed the best performance with an mAUC of 0.8359 and AUPRC of 0.5631, followed by the SAnD [14] model. Most recurrence-based methods presented relatively better performance compared to interpolation-based methods. GRU-D [7], which used both the missing indicators and time interval similar to the proposed method, demonstrated a competitive performance among the recurrence-based methods, which implies the necessity of utilizing the representations from the missing pattern. Furthermore, T-LSTM [8] exhibited the lowest performance, because it only considers the time intervals between the observations with a simple decay function.

These experimental results validated the efficacy of the proposed method that learns the multi-view representations of irregular time series data and their deep integration by the self-attention mechanism and further builds the auxiliary missing data imputer, showing its superior performance in various downstream tasks.

B. Ablation Study

According to Table III and Table IV, integrating the observation and sources of missingness yielded superior results than those obtained using only the observation, and the highest

TABLE VII
ABLATION STUDIES TO INVESTIGATE THE EFFECT OF USING THE IMPUTED DATA BY THE BRITS, AND AUXILIARY IMPUTER (IMP.) FOR THE PHYSIONET 2012 CHALLENGE DATASET

Task	Measure	Method		
		BRITS Imp.	Auxiliary Imp.	No Imp.
Mortality	AUC	0.8305±0.0156	0.8253±0.0235	0.8201±0.0218
	AUPRC	0.4765±0.0360	0.4742±0.0482	0.4698±0.0563
LOS	RMSE	11.1896±0.9764	12.0061±0.8311	11.5743±0.8937
	MAE	7.3364±0.1228	7.8168±0.1693	7.4762±0.1897

TABLE VIII
ABLATION STUDIES TO INVESTIGATE THE EFFECT OF USING THE IMPUTED DATA BY THE BRITS, AND AUXILIARY IMPUTER (IMP.) FOR THE MIMIC-III DATASET

Task	Measure	Method		
		BRITS Imp.	Auxiliary Imp.	No Imp.
Mortality	AUC	0.8598±0.0082	0.8562±0.0117	0.8534±0.0071
	AUPRC	0.3611±0.0217	0.3580±0.0153	0.3565±0.0133
LOS	RMSE	4.2052±0.0742	4.4119±0.0825	4.4336±0.0859
	MAE	2.1469±0.0764	2.1650±0.0878	2.0963±0.0859
Phenotyping	mAUC	0.8384±0.0043	0.8391±0.0031	0.8359±0.0034
	mAUPRC	0.5657±0.0076	0.5599±0.0042	0.5641±0.0085

TABLE IX
ABLATION STUDIES RELATED TO THE MISSING PATTERN GENERATION USING THE MISSING INDICATOR ($\mathbf{M}$) AND TIME INTERVAL ($\boldsymbol{\Delta}$) AS THE QUERY INPUT OF THE ATTENTION BLOCK FOR THE PHYSIONET 2012 CHALLENGE DATASET

Task	Measure	Method	
		$\mathbf{M}$ as query input	$\boldsymbol{\Delta}$ as query input
Mortality	AUC	0.8172±0.0325	**0.8201±0.0218**
	AUPRC	0.4597±0.0707	**0.4698±0.0563**
LOS	RMSE	12.0172±1.0211	**11.5743±0.8937**
	MAE	7.9694±0.4347	**7.4762±0.1897**

TABLE X
ABLATION STUDIES RELATED TO THE MISSING PATTERN GENERATION USING THE MISSING INDICATOR ($\mathbf{M}$) AND TIME INTERVAL ($\boldsymbol{\Delta}$) AS THE QUERY INPUT OF THE ATTENTION BLOCK FOR THE MIMIC-III DATASET

Task	Measure	Method	
		$\mathbf{M}$ as query input	$\boldsymbol{\Delta}$ as query input
Mortality	AUC	0.8475±0.0034	**0.8534±0.0071**
	AUPRC	0.3493±0.0238	**0.3565±0.0133**
LOS	RMSE	4.9251±0.0795	**4.4336±0.0859**
	MAE	2.5874±0.0801	**2.0963±0.0859**
Phenotyping	mAUC	0.8292±0.0031	**0.8359±0.0034**
	mAUPRC	0.5459±0.0060	**0.5641±0.0085**

performance over both the MIMIC-III and PhysioNet 2012 datasets in all downstream tasks corresponded to the integration for triple views. This finding demonstrated the effectiveness of incorporating the observation and sources of missingness. In the double view cases, the scenario of $(\mathbf{X} + \boldsymbol{\Delta})$ corresponded to a slightly higher performance in terms of the mortality and LOS prediction compared to that of the $(\mathbf{X} + \mathbf{M})$ case in both datasets for all downstream tasks.

To further examine the observation and sources of missingness, we conducted an additional ablation study to extract the missingness. Table IX and Table X showed the performance of the proposed model with $\boldsymbol{\Delta}$ or $\mathbf{M}$ as query input. In both datasets, the scenario of letting $\boldsymbol{\Delta}$ as a query achieved superior

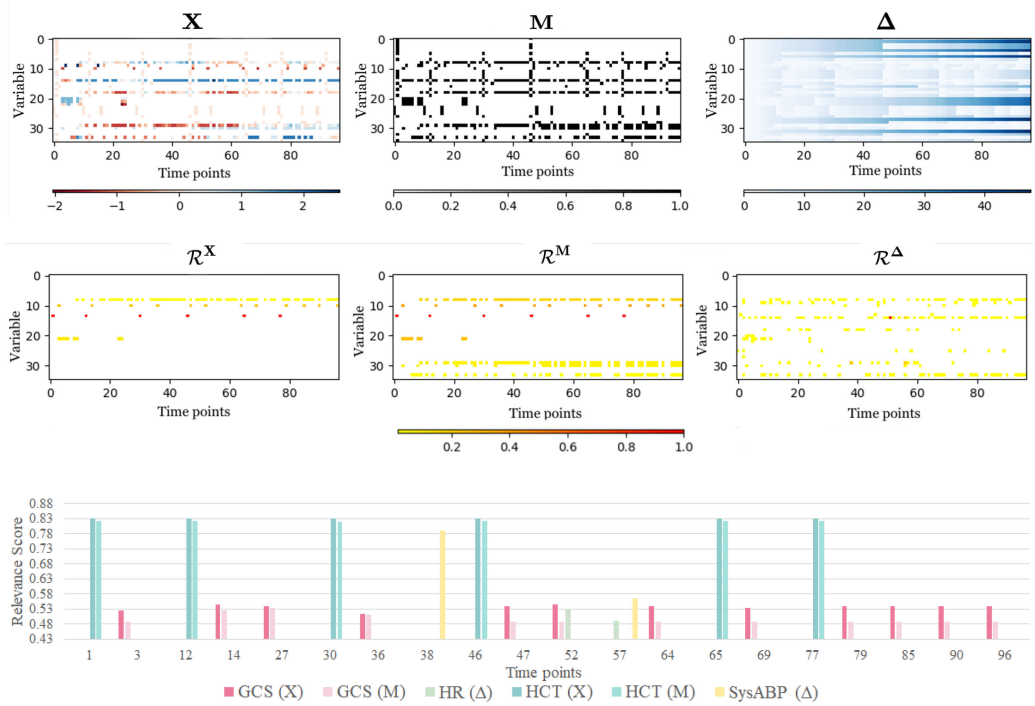

Fig. 3. Visualization of (top row) normalized raw data for observation ($\mathbf{X}$), missing indicator ($\mathbf{M}$), and time interval ($\mathbf{\Delta}$), (middle row) relevance score map for observation ($\mathcal{R}^{\mathbf{X}}$), missing indicator ($\mathcal{R}^{\mathbf{M}}$), and time interval ($\mathcal{R}^{\mathbf{\Delta}}$), and (bottom row) bar graph for changes of the highlighted relevance scores on their time points (GCS: Glasgow coma score, HR: heart rate, HCT: hematocrit, SysABP: systolic arterial blood pressure) of a patient with high mortality ($\hat{y} = 0.7922$). Here, we visualized only positive relevance scores for simplicity.

results. Combination of these results and those of previous ablation study, indicate that the time interval ($\mathbf{\Delta}$) provides key information regarding the missingness, although its importance weakens when it is used to generate the missing pattern with the missing indicator ($\mathbf{M}$).

The results in Table V and Table VI showed that the proposed model built with the attention-based imputer achieved a higher performance than that without the imputer in both datasets. The auxiliary imputer helps enhance the representational power of the interrelations among multi-view observations for the prediction task.

Furthermore, we compared our proposed approach to the case of using the imputed data by the BRITS imputer [17] as the SOTA imputation method and the auxiliary imputer for the PhysioNet and MIMIC-III dataset in the Table VII and Table VIII. The imputation (dotted line in Fig. 2) was also conducted in the test phase. The result obtained using the imputed data by the auxiliary imputer (Auxiliary Imp.) was superior to that of the proposed approach (No Imp.), and the (BRITS Imp.) achieved the highest performance. However, the imputation-based methods incurred significantly higher computing costs compared to the proposed method. Specifically, our method demanded 4.1091×10^{10} FLOPs (floating point operations per second), whereas the (BRITS Imp.) required 7.2526×10^{11} FLOPs due to their RNN-based imputation process. Considering the trade-off between the computing costs and performance, these results validated the effectiveness of the proposed method in downstream tasks.

Notably, these experimental results validated the efficacy of the proposed method that learns the multi-view representation of irregular time series data and their deep integration with the self-attention mechanism, and further builds an auxiliary missing data imputer, considering its superior performance in the downstream task. Furthermore, the ablation results indicated that the use of explicit imputation data helped enhance the prediction task performance to some extent compared to that when the imputation was not used, although the computational cost of the proposed method was significantly lower than that of the other approaches. Thus, the proposed approach in which the auxiliary imputer is used only in the training phase can adequately model irregular time series data.

Regarding the final representation, for which we used an average pooling, we compared with other strategies in Supplementary S2.

C. Layer-Wise Relevance Propagation (LRP) Analysis

In this study, to demonstrate the validity of our novel multi-view attention method in real-world risk prediction tasks, we applied the layer-wise relevance propagation (LRP) method [26] that has been widely employed to highlight the explainability of a learned model among attribution methods [27]. Specifically, we applied LRP-$\alpha\beta$ [26], [27], which treats positive and negative contributions in an asymmetric manner instead of vanilla LRP, and obtained the relevance scores.

We analyzed the in-hospital mortality prediction task over the PhysioNet 2012 challenge dataset. Specifically, we first calculated the relevance scores for each patient via the LRP and normalized them to a range of $(-1, 1)$. Next, we extracted the strongly highlighted variables and time points that showed high

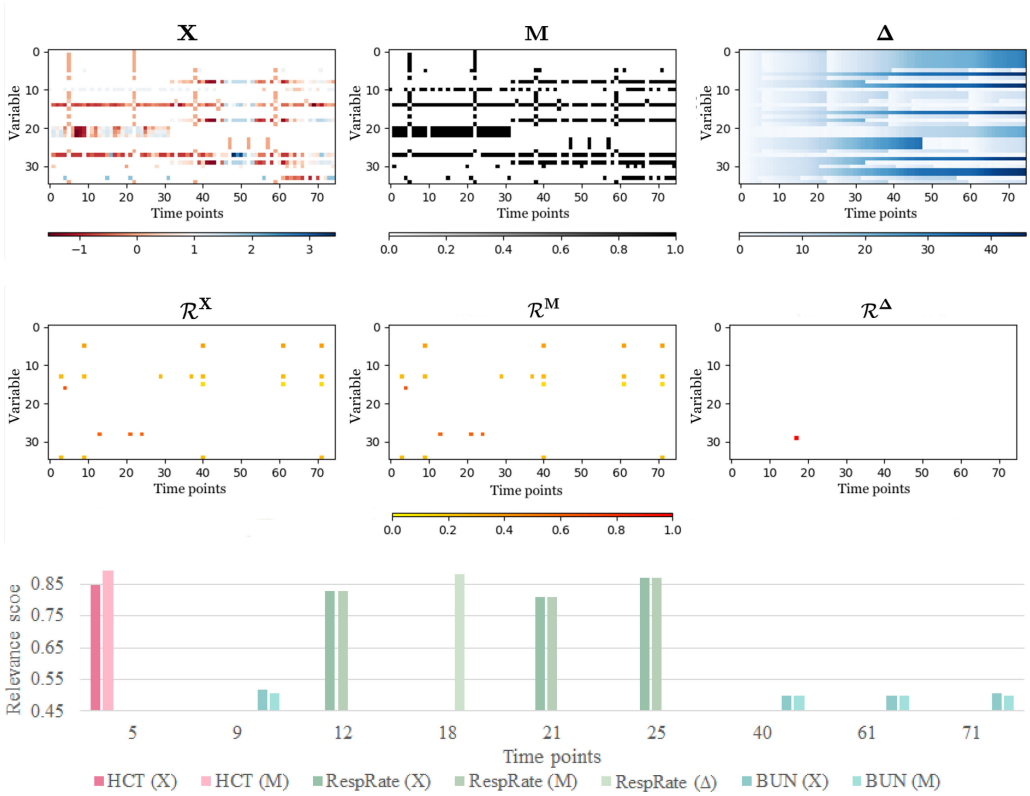

Fig. 4. Visualization of (top) normalized raw data for observation ($\mathbf{X}$), missing indicator ($\mathbf{M}$), and time interval ($\mathbf{\Delta}$), (middle) relevance score map for observation ($\mathcal{R}^{\mathbf{X}}$), missing indicator ($\mathcal{R}^{\mathbf{M}}$), and time interval ($\mathcal{R}^{\mathbf{\Delta}}$), and (bottom) bar graph for changes of the highlighted relevance scores on their time points (HCT: hematocrit, RespRate: respiration rate, BUN: blood urea nitrogen) of a patient with low mortality ($\hat{y} = 0.1542$). Here, we visualized only positive relevance scores for simplicity.

relevance scores, *i.e.*, those with scores greater than $\mu + 1.2\sigma$. Fig. 3 and Fig. 4 show the raw observation, missing indicators, and time interval ($\mathbf{X}, \mathbf{M}, \mathbf{\Delta}$), their corresponding relevance score map ($\mathcal{R}^{\mathbf{X}}, \mathcal{R}^{\mathbf{M}}, \mathcal{R}^{\mathbf{\Delta}}$), and the changes of highlighted relevance scores over the time points for the two patients with the highest and lowest probability among test samples with a ground truth label of 1, i.e. mortality.

1) Case I: Patient With the Highest Predicted Mortality Probability: Fig. 3 shows the input data and relevance scores of a patient with the highest predicted probability for in-hospital mortality. The relevance scores of all the input data showed relatively similar patterns, especially between $\mathcal{R}^{\mathbf{X}}$ and $\mathcal{R}^{\mathbf{M}}$. It can be inferred that the model effectively employed both the observation and missingness information during the decision process. The variables with high relevance score were Hematocrit (HCT), Glasgow coma score (GCS), systolic arterial blood pressure (SysABP), and heart rate (HR). HCT, which refers to the volume percentage of red blood cells in blood, is an indicator of heart failure in the ICU [28]. The raw signal at the time point of the high relevance scores showed a distinct change, affecting the increase in the mortality. The GCS is a clinical scale used to reliably measure a patient's level of consciousness after brain injury. The frequent measurement of the GCS and its changes indicate the severity of the patient's status, which is related to the increase in the mortality of the patient. Lastly, SysABP and HR presented high relevance scores over the time points at which they were frequently measured. The frequent measurement of both variables may indicate deterioration of the patient's condition, leading to the increased in-hospital mortality.

2) Case II: Patient With the Lowest Predicted Mortality Probability: Fig. 4 shows the input data and relevance scores of a patient with a mispredicted low probability, despite an actual ground truth label of 1. The variables with high relevance scores were respiration rate (RespRate), blood urea nitrogen (BUN), and HCT. RespRate indicates the number of breaths per minute, and thus, a rapid change in this value is considered abnormal. However, the relevance score of the patient was high when the RespRate exhibited minor changes. The BUN indicates the amount of urea nitrogen, a waste product of the digestion of protein, found in the blood. Although the BUN is the indicator for renal failure, it is not a distinct indicator of high in-hospital mortality.

The relevance scores of the patient with high mortality prediction were high in variables that are directly related to the mortality prediction, as well as in the time points involving distinct changes in the values. However, the patient with low mortality prediction showed a high relevance score for the variables or time points with low correlation to the in-hospital mortality. In addition, the experimental results for the high similarity of the relevance scores for the observations, missing indicators, and time interval data showed that the use of the representation of

the missingness and observation contributed to more accurate predictions.

VI. Conclusion

In this work, we proposed a method to directly learn the integrated representations of multi-view features from irregular multivariate time series data by using the self-attention mechanism without imputation. Specifically, we devised a novel multi-integration attention module (MIAM) to extract complex missing patterns by integrating missing indicators and time intervals, and further combine the observation and missing patterns in the representation space through a self-attention block. In addition, we built an attention-based decoder as a missing value imputer that helped empower the representation learning of the interrelations among multi-view observations; this imputer operated only in the training phase. We validated the effectiveness of our method in three downstream tasks, *i.e.*, prediction of in-hospital mortality, LOS prediction, and phenotyping, over the public MIMIC-III and PhysioNet challenge 2012 datasets and compared its performance with SOTA methods; the proposed method outperformed the SOTA methods. Furthermore, we identified the informative observations and time points by applying LRP to the learned model.

VI. Data Availability

MIMIC-III database analyzed in this study is available on PhysioNet repository. All the codes used for our experiments and analysis are open at https://github.com/ku-milab/MIAM.

References

[1] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," *IEEE J. Biomed.*, vol. 21, no. 1, pp. 22–30, Jan. 2017.

[2] Q. Suo *et al.*, "Personalized disease prediction using a CNN-based similarity learning method," in *Proc. Int. Conf. IEEE Bioinf. Biomed.*, 2017, pp. 811–816.

[3] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series," in *Proc. Mach. Learn. Health Conf.*, 2016, pp. 253–270.

[4] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1174–1182.

[5] S. N. Shukla and B. M. Marlin, "Interpolation-prediction networks for irregularly sampled time series," in *Proc. Int. Conf. Lern. Rep.*, 2019, pp. 1–14.

[6] M. Lechner and R. Hasani, "Learning long-term dependencies in irregularly-sampled time series," 2020, *arXiv:2006.04418*.

[7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, pp. 1–12, 2018.

[8] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 65–74.

[9] Y. Zhang, "ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4369–4375.

[10] Y. Zhang, X. Yang, J. Ivy, and M. Chi, "ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, vol. 34, pp. 930–937.

[11] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[12] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 787–795.

[13] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 743–752.

[14] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. Assoc. Adv. Artif. Intell.*, 2018, pp. 4091–4098.

[15] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt, "Set functions for time series," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4353–4363.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[17] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "BRITS: Bidirectional recurrent imputation for time series," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 6775–6785.

[18] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[19] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, "GP-VAE: Deep probabilistic time series imputation," in *Proc. Int. Conf. AISTATS*, 2020, pp. 1651–1661.

[20] E. Jun, A. W. Mulyadi, and H.-I. Suk, "Stochastic imputation and uncertainty-aware attention to EHR for mortality prediction," in *Proc. Int. Joint Conf. Neur. Netw.*, 2019, pp. 1–7.

[21] E. Jun, A. W. Mulyadi, J. Choi, and H. I. Suk, "Uncertainty-gated stochastic sequential model for EHR mortality prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4052–4062, Sep. 2021.

[22] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3391–3401.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171 4186.

[25] L. Liu *et al.*, "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Rep.*, 2019, pp. 1–13.

[26] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, 2015, Art. no. e0130140.

[27] Z. Wu and D. C. Ong, "On explaining your explanations of bert: An empirical study with sequence classification," *CoRR*, 2021, *arXiv:2101.00196*.

[28] B. D. Spiess *et al.*, "Hematocrit value on intensive care unit entry influences the frequency of Q-wave myocardial infarction after coronary artery bypass grafting," *J. Thoracic. Cardiovasc. Surg.*, vol. 116, pp. 460–467, 1998.