# UNCERTAINTY-AWARE VARIATIONAL-RECURRENT IMPUTATION NETWORK FOR CLINICAL TIME SERIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Electronic Health Records (EHR) comprise of longitudinal clinical observations portrayed with sparsity, irregularity, and high-dimensionality which become the major obstacles in drawing reliable downstream outcome. Despite greatly numbers of imputation methods are being proposed to tackle these issues, most of the existing methods ignore correlated features or temporal dynamics and entirely put aside the uncertainty. In particular, since the missing values estimates have the risk of being imprecise, it motivates us to pay attention to reliable and less certain information differently. In this work, we propose a novel variational-recurrent imputation network (V-RIN), which unified imputation and prediction network, by taking into account the correlated features, temporal dynamics, and further utilizing the uncertainty to alleviate the risk of biased missing values estimates. Specifically, we leverage the deep generative model to estimate the missing values based on the distribution among variables and a recurrent imputation network to exploit the temporal relations in conjunction with utilization of the uncertainty. We validated the effectiveness of our proposed model with publicly available real-world EHR dataset, PhysioNet Challenge 2012, and compared the results with other state-of-the-art competing methods in the literature.

## 1 INTRODUCTION

Electronic Health Records (EHR) store longitudinal data comprising of patient's clinical observations in the intensive care unit (ICU). Despite the surge of interest in clinical research on EHR, it still holds diverse challenging issues to be tackled with, such as high-dimensionality, temporality, sparsity, irregularity, and bias (Cheng et al., 2016; Lipton et al., 2016; Yadav et al., 2018; Shukla & Marlin, 2019). Specifically, sequences of the medical events are recorded irregularly in terms of variables and time, due to various reasons such as lack of collection or documentation, or even recording fault (Wells et al., 2013; Cheng et al., 2016). In fact, since it carries essential information regarding the patient's health status, improper handling of missing values might draw an unintentional bias (Wells et al., 2013; Beaulieu-Jones et al., 2017) yielding unreliable downstream analysis and verdict.

Complete-case analysis is one approach to draw the clinical outcome by disregarding the missing values and relying only on the observed values. However, excluding the missing data shows poor performance at high missing rates and also requires modeling separately for different dataset. In fact, the missing values and their patterns are correlated with the target labels (Che et al., 2018). Thus, we resort to the imputation approach to improve clinical outcomes prediction as the downstream task.

There exist numerous proposed strategies in imputing missing values in the literature. Brick & Kalton (1996) classified the imputation methods of being deterministic or stochastic in terms of the utilization of the randomness. While deterministic methods such as mean (Little & Rubin, 1987) and median filling (Acuña & Rodriguez, 2004) produced only one possible value, it is desirable to generate samples by considering the data distribution, thus leading to stochastic-based imputation methods. Moreover, since we are dealing with multivariate time series, an adequate imputation model should reflect several properties altogether, namely, 1) temporal relations, 2) correlations across variables, and additionally 3) offering a probabilistic interpretation for uncertainty estimation (Fortuin et al., 2019).

Recently, the rise of the deep learning models offers potential solutions in accommodating aforementioned conditions. Variational autoencoders (VAEs) (Kingma & Welling, 2014) and generative adversarial networks (GANs) (Goodfellow et al., 2014) exploited the latent distribution of high-dimensional incomplete data and generated comparable data points as the approximation estimates for the missing or corrupted values (Nazabal et al., 2018; Luo et al., 2018; Jun et al., 2019). However, even though these models employed the stochastic approach in inferring and generating samples, they scarcely utilized the uncertainty. In addition, such deep generative models are insufficient in estimating the missing values of multivariate time series, due to their nature of ignoring temporal relations between a span of time points. Hence, it requires additional approaches to model the temporal dynamics, such as Gaussian process (Fortuin et al., 2019) or recurrent neural network (RNNs) (Luo et al., 2018; Jun et al., 2019).

On the other hand, by the virtue of RNNs which have proved a remarkable performance in modeling the sequential data, we can estimate the complete data by taking into account the temporal characteristics. GRU-D (Che et al., 2018) proposed a modified gated-recurrent unit (GRU) cell to model missing patterns in the form of masking vector and temporal delay. Likewise, BRITS (Cao et al., 2018) modeled the temporal relations by bi-directional dynamics, and also considered features correlation by regression layers in estimating the missing values. However, they didn't take into account the uncertainty in estimating the missing values. That is, since the imputation estimates are not thoroughly accurate, we may introduce their fidelity score denoted by the uncertainty, which enhances the task performance by emphasizing the reliable or less uncertain information and vice versa (He, 2010; Gemmeke et al., 2010; Jun et al., 2019).

In this work, we define our primary task as prediction of in-hospital mortality on EHR data. However, since the data are characterized by sparse and irregularly-sampled, we devise an effective imputation model as the secondary problem but major concern in this work. We propose a novel variational-recurrent imputation network (V-RIN), which unified imputation and prediction network for multivariate time series EHR data, governing both correlations among variables and temporal relations. Specifically, given the sparse data, an inference network of VAE is employed to capture data distribution in the latent space. From this, we employ a generative network to obtain the reconstructed data as the imputation estimates for the missing values as well as the uncertainty indicating the imputation fidelity score. Then, we integrate the temporal and feature correlations into a combined vector and feed it into a novel *uncertainty-aware GRU* in the recurrent imputation network. Finally, we obtain the mortality prediction as a clinical verdict from the complete imputed data. In general, our main contributions in this paper are as follows:

- We estimate the missing values by utilizing deep generative model combined with recurrent imputation network to capture both features correlations and the temporal dynamics jointly, yielding the uncertainty.
- We effectively incorporate the uncertainty with the imputation estimates in our novel uncertainty-aware GRU cell for better prediction result.
- We evaluated the effectiveness of the proposed models by training the imputation and prediction networks jointly using the end-to-end manner, achieving the superior performance among other state-of-the-art competing methods on real-world multivariate time series EHR data.

## 2 Related Work

Imputation strategies were extensively devised to resolve the issue of sparse high-dimensional time series data by means of the statistics, machine learning, and deep learning methods. For instance, previous works exploited statistical attributes of observed data, such as mean (Little & Rubin, 1987) and median filling (Acuña & Rodriguez, 2004), which clearly ignored the temporal relations as well as the correlations among variables. From the machine learning approaches, expectation-maximization (EM) algorithm (Dempster et al., 1977), $k$-nearest neighbor (KNN) (Troyanskaya et al., 2001), principal component analysis (PCA) (Oba et al., 2003; Mohamed et al., 2009) were proposed by taking into account the relationships of the features either in the original or latent space. Furthermore, multiple imputation by chained equations (MICE) (White et al., 2011; Azur et al., 2011) introduced variability by means of repeating imputation process for multiple times. Yet, these methods ignore the temporal relations as the crucial attributes in time series modeling.

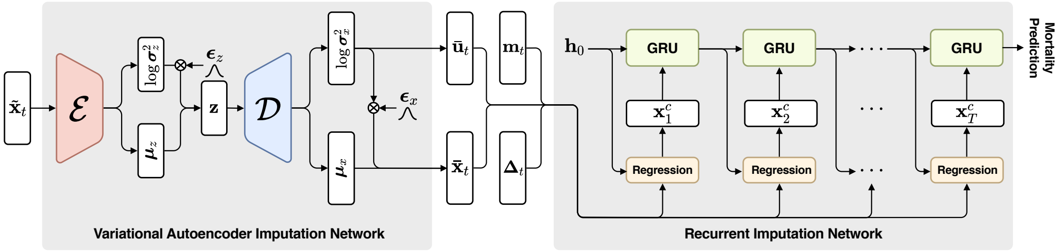


Figure 1: Architecture of the proposed model.

The deep learning-based imputation models are closely related to our proposed models. Nazabal et al. (2018) leveraged VAEs to generate stochastic imputation estimates by exploiting the distribution and correlations of features in the latent space. However, it ignores the temporal relations and the uncertainties as well. Recently, GP-VAE (Fortuin et al., 2019) were proposed to obtain the latent representation by means of VAEs and model the temporal dynamics in the latent space using Gaussian process. However, since the model is merely focused on the imputation task, they required a separate model for further downstream outcome.

To deal with the time series data, a series of RNNs-based imputation models were proposed. GRU-D (Che et al., 2018) took into account the temporal dynamics by incorporating the missing patterns, together with the mean imputation and forward filling with past values. Similarly, GRU-I (Luo et al., 2018) trained the RNNs using adversarial scheme of GANs as the stochastic approach. In the meantime, BRITS (Cao et al., 2018) were proposed to combine the feature correlations and temporal dynamics networks using bi-directional dynamics, which enhanced the accuracy by estimating missing values in both forward and backward directions. Likewise, M-RNN (Yoon et al., 2017) utilized bi-directional recurrent dynamics by operating interpolation (intra-stream) and imputation (inter-stream). Despite temporal dynamics are considered in their proposed models, yet the uncertainty for imputation was scarcely incorporated.

Our proposed model differs from the above models in ways of integrating imputation and prediction networks jointly. In particular, for the imputation network, we model both feature and temporal relations by means of deep generative model and recurrent imputation networks, respectively. Furthermore, we introduce the imputation fidelity of estimates as the uncertainty, compensating the potential impairment of imputation estimates. Specifically, it is noteworthy that our proposed model provides the reliable estimates, while giving the penalty to the unreliable ones determined by its uncertainties. Thereby, we expect to get better estimates of the missing values leading to a better prediction performance as a downstream task.

## 3 Proposed Method

Our architecture consists of two key networks: imputation and prediction network, as depicted in Fig. 1. The imputation network is devised on VAEs to capture the latent distribution of the sparse data by means of inference network (encoder $\mathcal{E}$). Then, the generative network (decoder $\mathcal{D}$) estimates reconstructed data distribution. We regard its mean as the imputation estimates, while exploiting the variance as the uncertainty to be further utilized in the recurrent imputation network for reliable prediction outcome.

Moreover, the succeeding recurrent imputation networks is built upon RNNs to model the temporal dynamics. In addition, for each time step, we use the regression layer to explore the feature correlation in imputing the missing values. Eventually, by unifying VAEs and RNNs systematically, we expect to acquire a more likely estimate by taking into account the features correlations, temporal relations over time as well as utilization of the uncertainty. We describe each of the networks more specifically in the following section after introducing the notations.

### 3.1 Data Representation

Given the multivariate time series EHR data of $N$ number of patients, a set of clinical observations and their corresponding label is denoted as $\{\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}\}_{n=1}^{N}$. For each patient, we denote $\mathbf{X}^{(n)} = [\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_t^{(n)}, \ldots, \mathbf{x}_T^{(n)}]^\top \in \mathbb{R}^{T\times D}$, where $T$ and $D$ represent time points and variables, respectively, $\mathbf{x}_t^{(n)} \in \mathbb{R}^D$ denotes all observed variables at $t$-th time point, and $x_t^{(n),d}$ is the $d$-th element of variables at time $t$. In addition, it has corresponding clinical label $\mathbf{y}^{(n)} \in \{0,1\}$ representing the in-hospital mortality which is a binary classification problem in our scenario. For the sake of clarity, hereafter we omit the superscript $(n)$.

To address the missing values, we introduce the masking matrix $\mathbf{M} \in \{0,1\}^{T\times D}$ indicating whether the values are observed or missing, and additionally define a new data representation $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_t, \ldots, \tilde{\mathbf{x}}_T]^\top \in \mathbb{R}^{T\times D}$, where we initialize the missing value with zero as follows:

$$m_t^d = \begin{cases} 1 & \text{if } x_t^d \text{ is observed,} \\ 0 & \text{otherwise} \end{cases}, \qquad \tilde{x}_t^d = \begin{cases} x_t^d & \text{if } m_t^d = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Besides, the time gap between the two observed values carries a piece of essential information. Thus, we further introduce time delay matrix $\mathbf{\Delta} \in \mathbb{R}^{T\times D}$, which is derived from $\mathbf{s} \in \mathbb{R}^T$, denoting the timestamp of the measurement. For the $t = 1$, we set $\mathbf{\Delta}_t = \mathbf{1}$. While for the rest $(t > 1)$, we set the time delay by referring to the masking matrix as follows:

$$\Delta_t^d = \begin{cases} s_t - s_{t-1} & \text{if } m_{t-1}^d = 1, \\ s_t - s_{t-1} + \Delta_{t-1}^d & \text{otherwise.} \end{cases}$$

### 3.2 VAE-based Imputation Network

Given the observations at each time point $\tilde{\mathbf{x}}_t$, we infer $\mathbf{z} \in \mathbb{R}^{k \ll D}$ as the latent representation by making use of the inference network, utilizing the true posterior distribution $p_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t)$. Intuitively, we assume that $\tilde{\mathbf{x}}_t$ is generated from some unobserved random variable $\mathbf{z}$ by some conditional distribution $p_\theta(\tilde{\mathbf{x}}_t|\mathbf{z})$, while $\mathbf{z}$ is generated from a prior distribution $p_\theta(\mathbf{z})$. Therefore, we define the marginal likelihood as $p_\theta(\tilde{\mathbf{x}}_t) = \int p_\theta(\tilde{\mathbf{x}}_t|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$. However, since it is intractable due to involvement of the true posterior $p_\theta(\mathbf{z}|\tilde{\mathbf{x}}_t)$, we approximate it with $q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t)$ using a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z^2)$, where the mean and log-variance are obtained such that:

$$\boldsymbol{\mu}_z = \mathcal{E}_\mu(\tilde{\mathbf{x}}_t; \phi), \quad \log \boldsymbol{\sigma}_z^2 = \mathcal{E}_\sigma(\tilde{\mathbf{x}}_t; \phi),$$

where $\mathcal{E}_{\{\mu,\sigma\}}$ denotes the inference network with parameter $\phi$. Furthermore, we apply the reparameterization trick proposed by Kingma & Welling (2014) as $\mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}_z$, where $\boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\odot$ denotes the element-wise multiplication, thus, making it possible to be differentiated and trained using standard gradient methods. Furthermore, given this latent vector $\mathbf{z}$, we estimate $p_\theta(\tilde{\mathbf{x}}_t|\mathbf{z})$ by means of the generative network $\mathcal{D}$ with parameter $\theta$ as :

$$\boldsymbol{\mu}_x = \mathcal{D}_\mu(\mathbf{z}; \theta), \quad \log \boldsymbol{\sigma}_x^2 = \mathcal{D}_\sigma(\mathbf{z}; \theta),$$

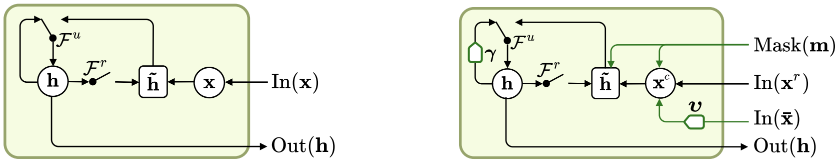
where $\boldsymbol{\mu}_x$ and $\boldsymbol{\sigma}_x^2$ denote the mean and variance of reconstructed data distribution, respectively. We apply another reparameterization trick in the data space to obtain $\hat{\mathbf{x}}_t = \boldsymbol{\mu}_x + \boldsymbol{\sigma}_x \odot \boldsymbol{\epsilon}_x$ with $\boldsymbol{\epsilon}_x \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We regard this reconstructed data as the estimates to the missing values and maintain the observed values in $\bar{\mathbf{x}}_t$ as follows:

$$\bar{\mathbf{x}}_t = \mathbf{m}_t \odot \tilde{\mathbf{x}}_t + (\mathbf{1} - \mathbf{m}_t) \odot \hat{\mathbf{x}}_t.$$

In the meantime, we regard the variance of reconstructed data as the uncertainty to be further utilized in the recurrent imputation process. For this purpose, we introduce an uncertainty matrix $\bar{\mathbf{U}} \in \mathbb{R}^{T\times D}$ with $\hat{\mathbf{\Sigma}}_x = [\text{diag}(\boldsymbol{\sigma}_{x,1}), \ldots, \text{diag}(\boldsymbol{\sigma}_{x,t}), \ldots, \text{diag}(\boldsymbol{\sigma}_{x,T})]^\top \in \mathbb{R}^{T\times D}$. We quantify this uncertainty as the fidelity score of the missing values etimates. In particular, we set the corresponding uncertainty as zero if the data is observed, indicating that we are confident with full trust to the observation, while set this as a value $\sigma_{x,t}^d$ if the corresponding value is missing as:

$$\bar{\mathbf{U}} = (\mathbf{1} - \mathbf{M}) \odot \hat{\mathbf{\Sigma}}_x$$

As a result of VAE-based imputation network, we obtain the set $\{\bar{\mathbf{X}}, \bar{\mathbf{U}}\}$ denoting the imputed values and its corresponding uncertainty, respectively.



(a) Vanilla GRU cell.

(b) Uncertainty-aware GRU cell.

Figure 2: Graphical illustrations of (a) vanilla GRU cell and (b) our modified GRU cell incorporating the uncertainty ($\mathcal{F}^u$: update gate, $\mathcal{F}^r$: reset gate).

### 3.3 Recurrent Imputation Network

The recurrent imputation network is based on RNNs, where we further model the temporal relations in the imputed data and exploit the uncertainties. While both GRU (Cho et al., 2014) or long-short term memory (LSTM) (Hochreiter & Schmidhuber, 1997) are feasible choices to be employed, inspired from the previous work of Che et al. (2018), we leverage the uncertainty-aware GRU cell to further consider uncertainty and the temporal decaying factor, which is depicted in Fig. 2b.

Specifically, at each time step $t$, we produce the uncertainty decay factor $\boldsymbol{\upsilon}_t$ in the Eq. (1) using negative exponentional rectifier to guarantee $\boldsymbol{\upsilon}_t \in (\mathbf{0}, \mathbf{1}]$, and further element-wise multiply this with $\bar{\mathbf{x}}_t$ to emphasizes the reliable estimates and give penalties to the uncertain ones expressed by the Eq. (2) resulting in $\mathbf{x}_t^{\upsilon}$.

$$\boldsymbol{\upsilon}_t = \exp\{-\max(0, \mathbf{W}_{\upsilon}\bar{\mathbf{u}}_t + \mathbf{b}_{\upsilon})\} \tag{1}$$

$$\mathbf{x}_t^{\upsilon} = \mathbf{m}_t \odot \bar{\mathbf{x}}_t + (\mathbf{1} - \mathbf{m}_t) \odot (\bar{\mathbf{x}}_t \odot \boldsymbol{\upsilon}_t) \tag{2}$$

By employing the GRU, we obtain the hidden state $\mathbf{h}$ as the comprehensive information compiled from the preceding sequences. Thus, given the previous hidden states $\mathbf{h}_{t-1}$, we produce the current complete observation estimates $\mathbf{x}_t^r$ through the following regression equation:

$$\mathbf{x}_t^r = \mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{b}_r \tag{3}$$

Hence, we have a pair of imputed values $\{\mathbf{x}_t^{\upsilon}, \mathbf{x}_t^r\}$ corresponding to missing values estimates based on features correlations and temporal relations, respectively. We then merge these information jointly to get combined vector $\mathbf{c}_t$ comprising both estimates:

$$\mathbf{c}_t = \mathbf{W}_c \left[\mathbf{x}_t^{\upsilon} \circ \mathbf{x}_t^r\right] + \mathbf{b}_c \tag{4}$$

where $\circ$ denotes a concatenation operation. Finally, we obtain the complete vector $\mathbf{x}_t^c$ by replacing the missing values with the combined estimates as follows:

$$\mathbf{x}_t^c = \mathbf{m}_t \odot \bar{\mathbf{x}}_t + (\mathbf{1} - \mathbf{m}_t) \odot \mathbf{c}_t \tag{5}$$

As time delay $\boldsymbol{\Delta}_t$ is essential element to capture temporal relations from the data (Che et al., 2018), we also introduce the temporal decay factor $\boldsymbol{\gamma}_t \in (\mathbf{0}, \mathbf{1}]$ in the Eq. (6) as

$$\boldsymbol{\gamma}_t = \exp\{-\max(0, \mathbf{W}_{\gamma}\boldsymbol{\Delta}_t + \mathbf{b}_{\gamma})\}. \tag{6}$$

We utilize this factor to control the influence of past observations embedded into hidden states using the form of $(\mathbf{h}_{t-1} \odot \boldsymbol{\gamma}_t)$. In addition to this, we concatenate the complete vector with corresponding mask, and then feed it into the uncertainty-aware GRU cell as illustrated in the Fig. 2b expressed as:

$$\mathbf{h}_t = \sigma(\mathbf{W}_h(\mathbf{h}_{t-1} \odot \gamma_t) + \mathbf{V}_h[\mathbf{x}_t^c \circ \mathbf{m}_t] + \mathbf{b}_h) \tag{7}$$

Lastly, to predict the in-hospital mortality as the clinical outcome, we utilize the last hidden state $\mathbf{h}_T$ to get the predicted label $\hat{y}$ such that:

$$\hat{y} = \sigma(\mathbf{W}_y\mathbf{h}_T + \mathbf{b}_y) \tag{8}$$

Note that $\mathbf{W}_{\{\upsilon,r,c,\gamma,h,y\}}$, $\mathbf{V}_h$ and $\mathbf{b}_{\{\upsilon,r,c,\gamma,h,y\}}\}$ are our learnable parameters in recurrent imputation network.

### 3.4 Learning

We specify the composite loss function comprising of the imputation and prediction loss function to tune all model parameters jointly, which are $\psi = \{\theta, \phi, \mathbf{W}_{\{\upsilon,r,c,\gamma,h,y\}}, \mathbf{V}_h, \mathbf{b}_{\{\upsilon,r,c,\gamma,h,y\}}\}$. By means of VAEs, we define the loss function $\mathcal{L}_{vae}$ to maximize the variational evidence lower bound (ELBO) that comprises of the reconstruction loss term and the Kullback-Leibler divergence. We add $\ell_1$-regularization to introduce the sparsity into the network with $\lambda_1$ as the hyperparameter. Moreover, we measure the difference between the observed data and the combined imputation estimates by the mean absolute error (MAE) as the $\mathcal{L}_{reg}$.

$$\mathcal{L}_{vae} = \sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t^{(n)})}[\log p_\theta(\tilde{\mathbf{x}}_t^{(n)}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t^{(n)})\|p_\theta(\mathbf{z})] + \lambda_1\|(\theta;\phi)\|_1 \tag{9}$$

$$\mathcal{L}_{reg} = \sum_{n=1}^{N}\sum_{t=1}^{T} \mathcal{L}_{MAE}(\bar{\mathbf{x}}_t^{(n)} \odot \mathbf{m}_t^{(n)}, \mathbf{c}_t^{(n)} \odot \mathbf{m}_t^{(n)}) \tag{10}$$

Furthermore, we define the binary cross-entropy loss function $\mathcal{L}_{pred}$ to evaluate the prediction of in-hospital mortality as follows:

$$\mathcal{L}_{pred} = -\sum_{n=1}^{N} y^{(n)} \log(\hat{y}^{(n)})$$

Finally, we define the overall loss function $\mathcal{L}_{total}$ as:

$$\mathcal{L}_{total} = \alpha\, \mathcal{L}_{vae} + \beta\, \mathcal{L}_{reg} + \mathcal{L}_{pred} + \lambda_2\|\psi\|_2^2 \tag{11}$$

with $\alpha$ and $\beta$ are the hyperparameters to represent the ratio between the $\mathcal{L}_{vae}$ and $\mathcal{L}_{reg}$, respectively, and $\lambda_2$ is the weight decay hyperparameter. Lastly, we use stochastic gradient decent in an end-to-end manner to optimize the model parameters during the training.

## 4 Experiments and Results

### 4.1 Dataset and Implementation Setup

We evaluated our proposed model on publicly available real-world EHR dataset, PhysioNet 2012 Challenge (Goldberger et al., 2000; Silva et al., 2012), which consists of 35 irregularly-sampled clinical variables (*i.e.* heart and respiration rate, blood pressure, *etc.*) from nearly 4,000 patients during first 48 hours of medical care in the ICU. We excluded 3 patients with no observation at all. We further sampled the observations hourly and take the last values in case of multiple measurements within this period, resulting data with mean missing rates of 80.51% and maximum of 99.84%. We predicted in-hospital mortality which are imbalanced with 554 positive mortality label (13.86%).

For the inference network of VAEs, we employed three layers of feedforward networks with hidden units of {64,24,10}, where 10 denotes the dimension of latent representation. The generative network has equal number of hidden units with those of inference network, but in the reverse order. We utilized Rectified Linear Unit (ReLU) as the non-linear activation function for each hidden layer. As for the recurrent imputation network, we used modified GRU with 64 hidden units. We trained the model using Adam optimizer with 200 epochs, 64 mini-batches and a learning rate of 0.0001. We set $\lambda_1$ and $\lambda_2$ equally with 0.0001. Finally, we reported the test result on in-hospital mortality prediction task from the 5-fold cross validation in terms of the average of Area Under the ROC curve (AUC) and Area Under Precision-Recall Curve (AUPRC). Additionally, to assess the effectiveness of our model in handling the imbalanced issue, we presented the balanced accuracy as well.

### 4.2 Comparative Models

We compared the performance of our proposed models with the closely-related state-of-the-art competing models in the literature:

- **M-RNN** (Yoon et al., 2017) exploited multi-directional RNNs which executing both interpretation and imputation to infer the missing data.

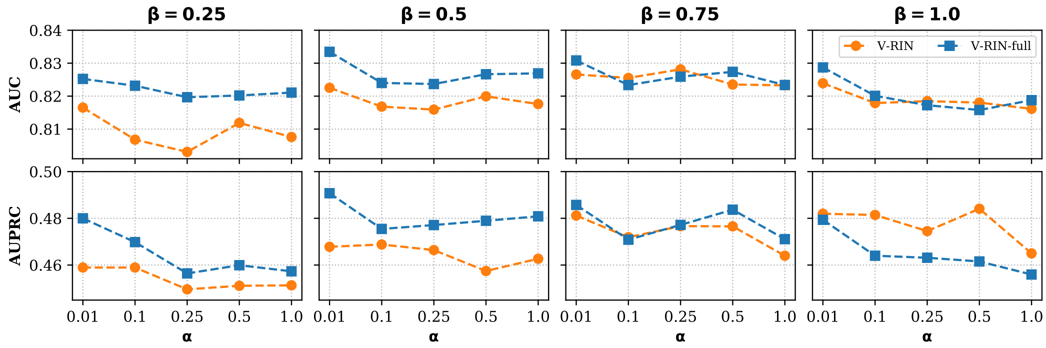


Figure 3: Result of ablation studies in terms of on the impact of imputation hyperparameters pair $\{\alpha, \beta\}$ in the classification task. V-RIN achieved the highest AUC score of $0.8281$ at $\{\alpha = 0.25, \beta = 0.75\}$, while V-RIN-full obtained the best average AUC of $0.8335$ using $\{\alpha = 0.001, \beta = 0.5\}$.

Table 1: Classification performance in the ablation studies (mean ± std).

| **Metric** | **VRNN (Chung et al., 2015)** | **VAE+RNN (Jun et al., 2019)** | **V-RIN (Ours)** | **V-RIN-full (Ours)** |
|---|---|---|---|---|
| AUC | $0.7926 \pm 0.0167$ | $0.8090 \pm 0.0167$ | $0.8281 \pm 0.0075$ | $\mathbf{0.8335 \pm 0.0106}$ |
| AUPRC | $0.4013 \pm 0.0352$ | $0.4414 \pm 0.0324$ | $0.4767 \pm 0.0257$ | $\mathbf{0.4907 \pm 0.0365}$ |

- **GRU-D** (Che et al., 2018) estimated the missing values by utilizing the missing patterns in forms of the masking and time decay factor.
- **GRU-I** (Luo et al., 2018) made use of adversarial scheme based on RNNs to consider both feature correlation and temporal dynamics altogether with its temporal decay as well.
- **BRITS** (Cao et al., 2018) utilized bi-directional dynamics in estimating the missing values based on features correlations and temporal relations. There are several variants of this model: **BRITS-I** utilized bi-directional dynamics considering only the temporal relations; **RITS** utilized unidirectional dynamics but use both feature and temporal correlations; while **RITS-I** utilized the unidirectional dynamics relying solely on temporal relations.
- **V-RIN** (Ours) is based on our proposed model except that we ignored the uncertainty. Specifically, we excluded the Eq. (1-2), and replaced $\mathbf{x}_t^{\upsilon}$ with $\bar{\mathbf{x}}_t$ in Eq. (4).
- **V-RIN-full** (Ours) executed all operations in the proposed model including feature-based correlations, temporal relations and the uncertainty to further mitigate the estimates bias.

## 4.3 Experimental Results

### 4.3.1 Ablation Studies

As part of the ablation studies, first we investigated the effect of varying the pair $\{\alpha, \beta\}$ as the hyperparameters of the imputation on both V-RIN and V-RIN-full model on the in-hospital mortality prediction task. We reflected these parameters as the ratio to weigh the imputation between feature and temporal relations in estimating the missing values in order to achieve the optimal performance. For each parameter, we defined a set of range values for these hyperparameters as $[0.01, 1.0]$ and presented the corresponding performances in the Fig 3. Both models were able to achieve high performance in terms of the average AUC score of $0.8281$ for V-RIN and $0.8335$ for V-RIN-full. V-RIN achieved its peak with setting $\{\alpha = 0.25,\ \beta = 0.75\}$, while V-RIN-full with $\{\alpha = 0.01,\ \beta = 0.5\}$. We interpreted these findings as by emphasizing more on the temporal relations than the features correlations in imputing the missing values, the model are able to obtain its best performance. However, once we tried to increase the $\beta$, we observed that the performance were degraded to some degree. Hence, it proved that both features and temporal are essential in estimating the missing values with some latent proportion.

Table 2: Performance on in-hospital mortality prediction task (mean ± std).

| **Models** | **AUC** | **AUPRC** | **Bal. ACC (%)** |
|---|---|---|---|
| M-RNN (Yoon et al., 2017) | $0.7718 \pm 0.0063$ | $0.3619 \pm 0.0199$ | $57.6114 \pm 2.2865$ |
| GRU-D (Che et al., 2018) | $0.8094 \pm 0.0142$ | $0.4571 \pm 0.0248$ | $57.9539 \pm 3.2072$ |
| GRU-I (Luo et al., 2018) | $0.7831 \pm 0.0205$ | $0.4029 \pm 0.0471$ | $58.3328 \pm 0.6374$ |
| RITS-I (Cao et al., 2018) | $0.8103 \pm 0.0137$ | $0.4511 \pm 0.0319$ | $61.5995 \pm 1.3904$ |
| RITS (Cao et al., 2018) | $0.8110 \pm 0.0129$ | $0.4558 \pm 0.0284$ | $60.3869 \pm 1.2256$ |
| BRITS-I (Cao et al., 2018) | $0.8184 \pm 0.0116$ | $0.4510 \pm 0.0351$ | $58.3711 \pm 3.4079$ |
| BRITS (Cao et al., 2018) | $0.8238 \pm 0.0100$ | $0.4782 \pm 0.0340$ | $59.4221 \pm 2.3565$ |
| **V-RIN (Ours)** | $0.8281 \pm 0.0075$ | $0.4767 \pm 0.0257$ | $62.0363 \pm 2.1644$ |
| **V-RIN-full (Ours)** | $\mathbf{0.8335 \pm 0.0106}$ | $\mathbf{0.4907 \pm 0.0365}$ | $\mathbf{63.1127 \pm 3.0619}$ |

Furthermore, Table 1 presented the comparison of our model with closely related models such as VRNN (Chung et al., 2015) which integrates VAEs for each time steps of RNNs. However, we observed that the performance is considered as low in terms of AUC and AUPRC as well. In addition, we compared also with VAEs followed by RNNs (VAE+RNN) without incorporating neither the temporal decay factor nor the uncertainty (Jun et al., 2019). We noticed a performance improvement in VAE+RNN which executes the imputation process by firstly exploiting the features correlations followed by temporal dynamics in exact order. Furthermore, by introducing the temporal decay in V-RIN, it helped a lot for the model to learn the temporal dynamics effectively resulting better AUC and AUPRC. Finally, once we introduced the uncertainty which is incorporated in the imputation network of V-RIN-full, we observed a significant enhancement of both AUC and AUPRC. This is undeniable evidence of the advantage of utilizing the uncertainty in further downstream task.

### 4.3.2 Prediction Result Analysis

We presented the experimental result of the in-hospital mortality prediction in comparison with other competing methods in terms of average AUC, AUPRC and balanced accuracy in Table 2. In practice, we removed 10% of the observed data randomly to make a fair comparison with BRITS model variants. Our V-RIN model is directly comparable to other competing models in ways that it estimates the missing values without incorporating the uncertainty. It achieved better performance in terms of AUC and balanced accuracy, and slightly comparable to BRITS in terms of AUPRC. However, we note that BRITS-I and BRITS utilized bi-directional dynamics to estimate the missing values achieving relatively higher performance. Although M-RNN employed similar strategies using bi-directional dynamics, it struggles to perform the task properly. Furthermore, our V-RIN is closely related to GRU-D, GRU-I, RITS-I, and RITS in ways of exploiting the temporal decay factor. However, V-RIN outperformed all aforementioned models indicating the effectiveness of missing values estimation using deep generative models by means of VAEs. Meanwhile, GRU-I which also makes use of deep generative model using adversarial strategies performed inferior compared to our model. This might be due to the fact that they employed imputation and prediction model separately. Ultimately, we obtained the highest overall performance results with the proposed V-RIN-full including the average balanced accuracy, indicating the effectiveness of the model in handling the imbalance issue which is non-trivial. Thereby, these findings reassure that the utilization of the uncertainty is truly beneficial in estimating the missing values. Hence, our model were able to achieve reliable downstream task and outperformed all comparative models.

## 5 Conclusions

In this paper, we proposed a novel unified framework comprising of imputation and prediction network for sparse high-dimensional multivariate time series. It combined deep generative model with recurrent model to capture features correlations and temporal relations in estimating the missing values and yielding uncertainty. We utilized the uncertainties as the fidelity of our estimation and incorporated them for clinical outcome prediction. We evaluated the effectiveness of proposed model with PhysioNet 2012 Challenge dataset as the real-world EHR multivariate time series data, proving the superiority of our model in the in-mortality prediction task, compared to other state-of-the-art comparative models in the literature.

## REFERENCES

Edgar Acuña and Caroline Rodriguez. *The Treatment of Missing Values and its Effect on Classifier Accuracy*. Springer Berlin Heidelberg, 2004.

Melissa J. Azur, Elizabeth Stuart, Constantine Frangakis, and Philip Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 3 2011. ISSN 1049-8931. doi: 10.1002/mpr.329.

Brett K. Beaulieu-Jones, Jason H. Moore, and et al. Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 22, pp. 207–218, 2017.

JM Brick and G. Kalton. Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238, 1996. doi: 10.1177/096228029600500302. URL https://doi.org/10.1177/096228029600500302. PMID: 8931194.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6775–6785. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7911-brits-bidirectional-recurrent-imputation-for-time-series.pdf.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. In *Scientific Reports*, volume 8, 2018. doi: 10.1038/s41598-018-24271-9. URL https://doi.org/10.1038/s41598-018-24271-9.

Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 432–440, 2016.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2980–2988. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5653-a-recurrent-latent-variable-model-for-sequential-data.pdf.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Vincent Fortuin, Gunnar Rätsch, and Stephan Mandt. Multivariate time series imputation with variational autoencoders. *arXiv preprint arXiv:1907.04155*, 2019.

Jort Gemmeke, Ulpu Remes, and Kalle J Palomäki. Observation uncertainty measures for sparse imputation. In *Proceedings Interspeech*, pp. 2262–2265, 2010.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Yulei He. Missing data analysis using multiple imputation. *Circulation: Cardiovascular Quality and Outcomes*, 3(1):98–105, 2010. doi: 10.1161/CIRCOUTCOMES.109.875658. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.109.875658.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Eunji Jun, Ahmad Wisnu Mulyadi, and Heung-Il Suk. Stochastic imputation and uncertainty-aware attention to EHR for mortality prediction. *International Joint Conference on Neural Networks (IJCNN)*, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Zachary C Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 253–270, Children's Hospital LA, Los Angeles, CA, USA, 18–19 Aug 2016. PMLR. URL http://proceedings.mlr.press/v56/Lipton16.html.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 1987.

Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, and Yuan xiaojie. Multivariate time series imputation with generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1596–1607. Curran Associates, Inc., 2018.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 21*, pp. 1089–1096. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3532-bayesian-exponential-family-pca.pdf.

Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using VAEs. *arXiv Preprint arXiv:1807.03653*, 2018.

Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 11 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg287. URL https://doi.org/10.1093/bioinformatics/btg287.

Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1efr3C9Ym.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G. Mark. Predicting in-hospital mortality of ICU patients: The physionet/computing in cardiology challenge 2012. *Computing in Cardiology 2012*, 39, 2012.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 06 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520. URL https://doi.org/10.1093/bioinformatics/17.6.520.

Brian J. Wells, Kevin M. Chagin, Amy S. Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. In *eGEMS: The Journal of Electronic Health Data and Methods*, volume 1, pp. 1035, 2013. doi: 10.13063/2327-9214.1035. URL `http://doi.org/10.13063/2327-9214.1035`.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011. doi: 10.1002/sim.4067. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4067`.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs): A survey. *ACM Comput. Surv.*, 50(6):85:1–85:40, January 2018. ISSN 0360-0300. doi: 10.1145/3127881. URL `http://doi.acm.org/10.1145/3127881`.

Jinsung Yoon, William R. Zame, and M. van der Schaar. Multi-directional recurrent neural networks: A novel method for estimating missing data. In *International Conference on Machine Learning (ICML) Time Series Workshop*, 2017.