IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Stochastic Imputation and Uncertainty-Aware Attention to EHR for Mortality Prediction

Eunji Jun*, Ahmad Wisnu Mulyadi*, and Heung-Il Suk†
Department of Brain and Cognitive Engineering, *Korea University*
Seoul 02841, Republic of Korea
Email: {ejjun92, wisnumulyadi, hisuk}@korea.ac.kr

***Abstract*—Electronic health records (EHR) have become an important source of a patient data but characterized by a variety of missing values. Using the variational inference of Bayesian framework, variational autoencoder (VAE), a deep generative model, has been shown to be efficient and accurate to capture the latent structure of complex high-dimensional data. Recently, it has been used for missing data imputation. In this paper, we propose a general framework that incorporates effective missing data imputation using VAE and multivariate time series prediction. We utilize the uncertainty obtained from the generative network of the VAE and employ uncertainty-aware attention in imputing the missing values. We evaluated the performance of our architecture on real-world clinical dataset (MIMIC-III) for in-hospital mortality prediction task. Our results showed higher performance than other competing methods in mortality prediction task.**

***Index Terms*—Missing data imputation, Electronic health records, Bayesian framework, Deep learning**

I. Introduction

Electronic health records (EHR) have become a vital source of patient care since it stores the longitudinal record of clinic-related events, which are carried for each patient's encounters. They consist of heterogeneous data including demographic information, diagnoses, laboratory results, medication, and other clinic-related information. In most of the cases, these types of information are not entirely observed due to various reasons such as a lack of collection or documentation [2]. While these missing values and patterns actually provide rich information about the target label [17], it is challenging to do the analytical procedure using data with missing values. In the meantime, inadequate handling of missing data may result in biased estimation in further data analysis. Hence, developing an appropriate imputation model for missing values is a crucial step to be conducted before further steps.

There exist many imputation techniques in the literature such as statistical methods (*e.g.*, mean substitution, hot-deck and multiple imputation) and machine-learning methods (*e.g.*, k-nearest neighbor (KNN), singular value decomposition (SVD), principal component analysis (PCA), multi-layer perceptron (MLP)). However, they don't take into account the distribution of data or relationship between the variables. In addition, because the imputed values are completely determined by observations, they can be underestimated or overestimated.

To address this limitation, we exploit a generative model, which allows us to learn any kind of data distribution in an unsupervised manner, and infer the conditional relationship between observable and latent variables. By estimating the data distribution of the training set, we generate new data points with some variations. Here, we use a *Bayesian framework*, which is concrete and robust for building a generative model.

In a Bayesian framework, a joint distribution of latent and observable variables is defined, and the data distribution is learned to account for observed data with the highest probability. However, learning the exact distribution implicitly or explicitly is challenging since inference of posterior is intractable in most cases. In order to approximate the distribution, we especially exploit variational autoencoder (VAE) [5], one of the deep generative models, which is based on the variational inference. It has shown to achieve impressive performance in capturing the latent representation of the complex high dimensional data in an unsupervised manner [1], [5]. To learn a function which can approximate the model distribution to the actual distribution, the neural network is used.

Based on the learned relationship between latent variables and data, we can impute the missing data using the reconstructed data obtained by a generative network in our VAE. In addition, in this paper, we introduce the uncertainty for missing value estimation from the variance as the output of generative network. The uncertainty estimates account for increased variability from missing values. We can reduce the risk of biased estimation by reflecting the uncertainty in the predictive model.

In this work, we propose a novel Bayesian framework for missing value imputation in ICU patient records of EHR and in-hospital mortality prediction by reflecting the uncertainty into the predictive model. The major contributions of the proposed method are as follows:

- We utilize the Bayesian framework for missing values imputation method so that we can describe the observations from latent space in a probabilistic manner.
- To the best of our knowledge, this is the first work to introduce the '*uncertainty*' for the imputation estimates of missing values.
- We trained the imputation model coupled with the predictive model in an end-to-end manner, effectively exploring

†Corresponding author
*Equally contributed

978-1-7281-2009-6/$31.00 ©2019 IEEE

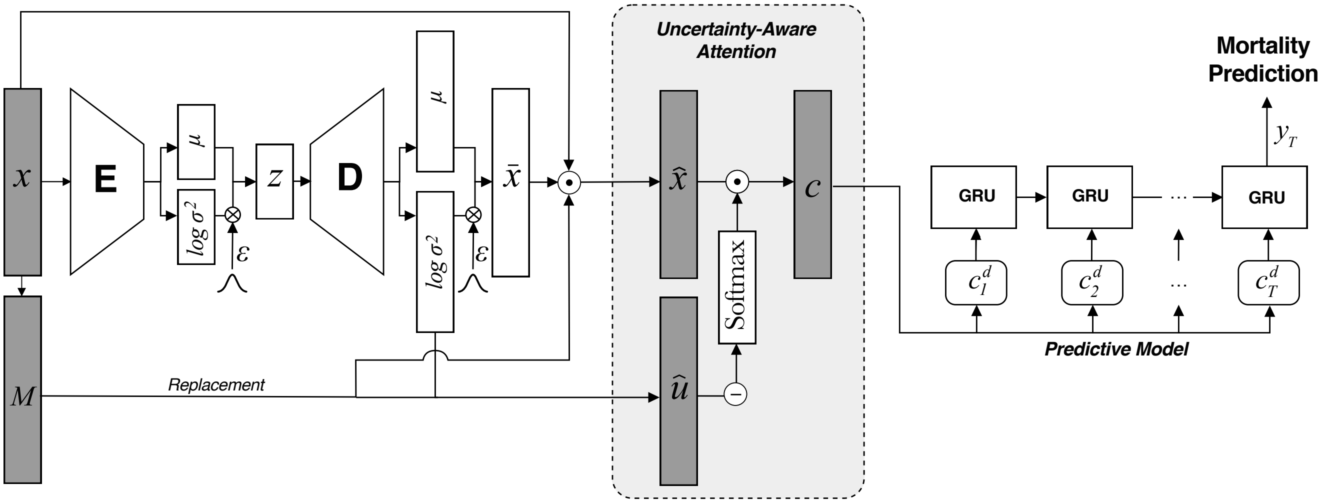


Fig. 1. Architecture of the proposed model.

the missing patterns and leading to achieve the optimal performances.

Our proposed method was empirically validated using the publicly available real world EHR dataset, the Medical Information Mart for Intensive Care III (MIMIC-III), showing better performances than other competitive methods.

II. Related works

A. Missing value imputation

There are various approaches to impute missing values in observations. Zero or mean substitution, KNN, and soft imputation, multivariate imputation by chained equations (MICE) [10] are widely used in practice. Despite their widespread use, it is noted that some of the methods are not able to capture the hidden relationship nor the distributions of the variables [3], [12], [17].

There were efforts to employ methods incorporating the relationship and distribution of the data by employing generative models. Autoencoder (AE) [3] and generative adversarial networks (GAN) [8] are the deep learning-based generative models which were used recently in order to tackle missing data imputation. As an AE can be trained to learn the relation among variables by transforming the data into its representation in a latent space, [3] employed undercomplete autoencoders in order to impute the missing data. However, as to further exploit the latent distribution of the data, we take the probabilistic approach in the latent space. As for another generative model, generative adversarial imputation nets (GAIN) is a model based on GAN that introduces a generator for imputing data to fool a discriminator that has to distinguish whether the data was observed or generated one [8]. Their approach to capture the uncertainty is to make a multiple draws for multiple imputations. This method is different from our proposed method in a way that we introduce the uncertainty by further incorporating the variance from the VAE's decoder.

B. Predictive model

Recurrent neural network (RNN)-based models have proved to achieve high performance in modeling the time series data and capturing temporal and non-linearity relations among them. Given the EHR data, [7] employed gated recurrent units (GRU), one of the variants of RNN, to predict heart failure. [16] used the long short-term memory (LSTM) to tackle the problem of diagnoses as the multi-label classification. When comparing to traditional machine learning methods such as logistic regression (LR), SVM, KNN and MLP, they exhibited the superiority of their method in performance.

As we mainly focus on the problem of missing value imputation for the high dimensional time series EHR data, to evaluate our proposed imputation method, we employ the existing RNN model, GRU, for in-hospital mortality prediction, and compare the result with traditional machine learning model as non-RNN model (*e.g.*, LR).

III. Proposed Model

We define the problem of predicting in-hospital mortality as a multivariate time series classification. Our goal is to predict whether the patient admitted to the hospital dies during the hospital stay after admission to an ICU. The overall architecture of the proposed model is shown in Fig. 1. The data containing the missing values are replaced with zero values and fed into the VAE, where we use a masking vector to preserve the missingness. From the model, we can obtain the estimated imputation values along with their uncertainty as well. By using this imputed data as the input to the RNN classifier to predict the mortality prediction, we train the proposed model using end-to-end fashion to get better estimation for the missing values imputation.

In this section, we first describe the overall notation for the structure of sequential EHR data, and then introduce our framework that estimates the missing values in the data using Bayesian approach followed by a predictive model.

A. Data representation

We consider a multivariate time series with T time points as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)^\top \in \mathbb{R}^{T \times D}$, where $t \in \{1, 2, ..., T\}$ and $\mathbf{x}_t \in \mathbb{R}^D$ represents the t-th observation of D variables. x_t^d denotes the element of the d-th variable in $\mathbf{x}_t$. We introduce the masking vector, which has the same size of time series data, $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_T)^\top \in \mathbb{R}^{T \times D}$, where $\mathbf{m}_t \in \{0, 1\}^D$ denotes which variables are observed at time t. Specifically, we have $m_t^d = 1$ if x_t^d is observed, otherwise, $m_t^d = 0$.

For the data including missing values, we define a new multivariate time series $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T)^\top \in \mathbb{R}^{T \times D}$ as follows:

$$\tilde{x}_t^d = \begin{cases} x_t^d, & \text{if } m_t^d = 1 \\ *, & \text{otherwise} \end{cases} \tag{1}$$

where * is an unobserved value to be filled through an imputation technique. Given time series dataset $\mathcal{D} = \{(\tilde{\mathbf{X}}^{(n)}, \mathbf{M}^{(n)})\}_{n=1}^N$, we predict the mortality labels $y^{(n)} \in \{0, 1\}$ in the binary classification problem.

B. Missing value imputation using Bayesian approach

In our approach, we initially replace *, the unobserved values in $\tilde{\mathbf{X}}$, with zeros and utilize the Bayesian approach in order to impute them. We assume that for each timestamp, an individual observation, $\tilde{\mathbf{x}}_t$, is generated by underlying distribution $p(\tilde{\mathbf{x}}_t)$. As such, the latent variable $\mathbf{z}$, is produced by some distribution $p(\mathbf{z})$, where we can interpret it as the hidden health status of the patient.

The joint distribution, $p(\tilde{\mathbf{x}}_t, \mathbf{z})$ can be represented as:

$$p(\tilde{\mathbf{x}}_t, \mathbf{z}) = p(\mathbf{z}) p(\tilde{\mathbf{x}}_t | \mathbf{z}). \tag{2}$$

In Equation (2), the joint distribution can be generated by sampling from the prior, the distribution of $\mathbf{z}$, and the likelihood $p(\tilde{\mathbf{x}}_t|\mathbf{z})$, the distribution of $\tilde{\mathbf{x}}_t$ given $\mathbf{z}$. The likelihood $p(\tilde{\mathbf{x}}_t|\mathbf{z})$ can be estimated from a *generative network*, and the prior $p(\mathbf{z})$ is typically selected to follow the multi-dimensional normal distribution with zero mean and unit variance [5]. The marginal likelihood of the model, $p(\tilde{\mathbf{x}}_t)$ can be obtained by integration over Equation (2) as follows:

$$p(\tilde{\mathbf{x}}_t) = \int p(\tilde{\mathbf{x}}_t, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{z}) p(\tilde{\mathbf{x}}_t|\mathbf{z}) d\mathbf{z}. \tag{3}$$

From Bayes' rule, the marginal likelihood $p(\tilde{\mathbf{x}}_t)$ is obtained by the integration of the joint distribution $p(\tilde{\mathbf{x}}_t, \mathbf{z})$ for $\mathbf{z}$. It includes the posterior distribution $p(\mathbf{z}|\tilde{\mathbf{x}}_t)$, which is the distribution of the latent variable $\mathbf{z}$ given observation data $\tilde{\mathbf{x}}_t$, $p(\mathbf{z}|\tilde{\mathbf{x}}_t)$ as follows:

$$p(\mathbf{z}|\tilde{\mathbf{x}}_t) = \frac{p(\tilde{\mathbf{x}}_t|\mathbf{z}) p(\mathbf{z})}{p(\tilde{\mathbf{x}}_t)}. \tag{4}$$

However, the posterior $p(\mathbf{z}|\tilde{\mathbf{x}}_t)$ is analytically intractable because the calculation of the evidence term $p(\tilde{\mathbf{x}}_t)$ is not possible. To tackle such intractable problem, the approximation method is required to infer $p(\mathbf{z}|\tilde{\mathbf{x}}_t)$. To this end, we use a stochastic variational Bayesian inference, one of the widely used approximation methods, where the posterior is approximated by $q(\mathbf{z}|\tilde{\mathbf{x}}_t)$.

In specific, we can effectively handle it by means of the VAE, a neural network that consists of *encoder* corresponding to inference network with parameters ϕ and *decoder* corresponding to generative network with parameters θ. By changing the parameters ϕ and θ, the best representation of the data $\tilde{\mathbf{x}}_t$ is obtained so that the logarithm of the marginal likelihood $\ln p_\theta(\tilde{\mathbf{x}}_t)$ is maximized. The log-evidence can be expressed in terms of the evidence lower bound (ELBO) as follows:

$$\ln p(\tilde{\mathbf{x}}_t) = ELBO + \mathcal{D_{KL}}[q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t) || p(\mathbf{z})]. \tag{5}$$

where $\mathcal{D_{KL}}$ denotes a Kullback-Liebler (KL) divergence that represents the agreement between two distributions, $q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t)$ and $p(\mathbf{z})$. The KL divergence is zero if the two distributions are identical, and positive for non-identical distributions. Since the KL divergence is always non-negative, the log-evidence is always greater than or equal to the ELBO as:

$$\ln p(\tilde{\mathbf{x}}_t) \geq ELBO. \tag{6}$$

Hence, by finding the parameters ϕ and θ that maximize the ELBO, the log-evidence can be derived towards its maximum value, defined as:

$$ELBO = \mathbb{E}_{q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t)}[\ln p(\tilde{\mathbf{x}}_t)] - \mathcal{D_{KL}}[q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t) || p(\mathbf{z})]. \tag{7}$$

In equation (7), the first term corresponds to reconstruction loss, which is an expectation for the log-likelihood of the generative network given the output of the inference network. We can train the network parameters ϕ and θ via stochastic gradient descent and back-propagation. With the trained VAEs, we sample $\epsilon \sim \mathcal{N}(0, I)$, and generate $\mathbf{z}$ given $\tilde{\mathbf{x}}_t$ by sampling from the latent variable distribution, $\mathcal{N}(\mu_\phi(\tilde{\mathbf{x}}_t), \Sigma_\phi(\tilde{\mathbf{x}}_t))$, which is the output of the inference network. Here, we use reparameterization trick to compute $\mu_\phi(\tilde{\mathbf{x}}_t) + \Sigma_\phi(\tilde{\mathbf{x}}_t) * \mathcal{N}(0, I)$, which makes the network differentiable.

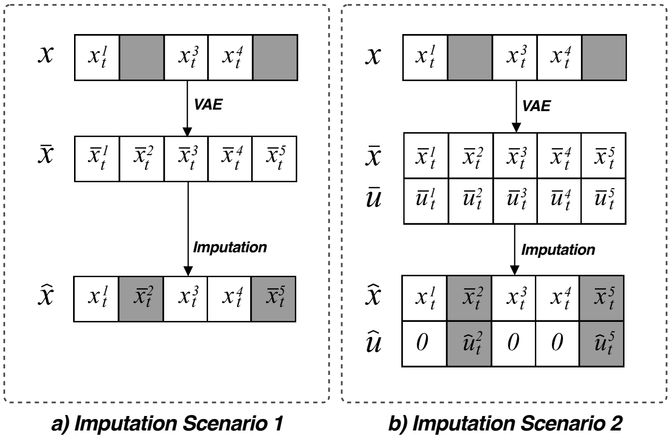
From the generative network, we generate $\bar{\mathbf{x}}_t$, given $\mathbf{z}$, by sampling from the reconstructed data distribution, $\mathcal{N}\left(\mu_\theta(\mathbf{z}), \text{diag}[\hat{\sigma}_t^i]_{i=1}^D\right)$. We replace the missing values with the reconstructed values, leaving the observed values unchanged. The reconstruction error is computed only for the observed values. The resulting imputed data, $\hat{\mathbf{X}}$, has the following form:

$$\hat{\mathbf{X}} = \mathbf{M} \odot \tilde{\mathbf{X}} + (\mathbf{1} - \mathbf{M}) \odot \bar{\mathbf{X}}. \tag{8}$$

where $\odot$ denotes element-wise multiplication.

From $\text{diag}[\hat{\sigma}_t^i]_{i=1}^D$ of the reconstruction distribution, *uncertainty* of the imputed values for each time step, $\bar{\mathbf{u}}_t$, can be estimated with the same size of the data $\mathbf{x}_t$, leading an uncertainty matrix, $\bar{\mathbf{U}}$ for all time steps. For the points that exist the observations, we replace $\bar{u}_t^d$ with zeros to guarantee that the actual measurements are not uncertain, resulting in $\hat{\mathbf{U}}$ with the following condition:

$$\hat{u}_t^d = \begin{cases} 0, & \text{if } m_t^d = 1 \\ \bar{u}_t^d, & \text{otherwise} \end{cases} \tag{9}$$



a) Imputation Scenario 1 ***b) Imputation Scenario 2***

Fig. 2. Two VAE-based imputation scenarios. a) Scenario 1 uses only the reconstructed data from the VAE while b) Scenario 2 incorporates the uncertainty obtained from the decoder of the VAE.

C. Uncertainty-aware mortality prediction

For time-series classification, we use GRU [9], which is widely used for strong prediction performance as well as the ability to capture long-term temporal dependencies. In this work, we further investigate the application of uncertainty for missing value imputation into the predictive model. As illustrated in Fig. 2, we come up with two VAE-based imputation scenarios. In the first scenario, we only use $\hat{\mathbf{x}}_t$ for each timestep as the estimated imputation results. In contrast, in the scenario 2, we further incorporate the uncertainty $\hat{\mathbf{u}}_t$ which we obtained from the decoder of the VAE. Therefore, the input to GRU will be $\hat{\mathbf{x}}_t$ and also the uncertainty $\hat{\mathbf{u}}_t$.

For each timestamp, the imputed data is combined with the uncertainty vector in a nonlinear manner before entering the predictive model. We take the softmax at the negative value of the uncertainty, and element-wise multiply its output with the imputed data, where we call the resulting vector as a combination vector, $\mathbf{c}_t \in \mathbb{R}^D$, at different time steps $t = 1, 2, ..., T$. The basic idea underlying the above approach is to reduce the impact of imputed data on predictive model for highly uncertain values. That is, the uncertainty estimates can be regarded as the fidelity of the imputation values. Equation (10) is defined to obtain a combination vector $\mathbf{c}_t \in \mathbb{R}^D$ for each time step. Thus, the combination vector is fed into $\mathbf{h}_t$, a hidden state of GRU.

$$\mathbf{c}_t = \hat{\mathbf{x}}_t \odot \left(\frac{\exp(-\hat{\mathbf{u}}_t)}{\sum_{d=1}^{D} \exp(-\hat{u}_t^d)} \right). \tag{10}$$

GRU learns a series of hidden state vectors $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T = \text{RNN}(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T)$. Each hidden state vector, $\mathbf{h}_t \in \mathbb{R}^p$, can be represented by $\mathbf{h}_t = f(\mathbf{c}_t; \mathbf{w}, b)$, where $f(\mathbf{c}_t; \mathbf{w}, b)$ is a function that maps input combination sequences to a single output, parameterized by $\mathbf{w} \in \mathbb{R}^{D \times M}$, $b \in \mathbb{R}^M$ that are shared parameters across all time steps. We use the last hidden vector $\mathbf{h}_T$ to predict an outcome label y.

To learn the model parameters, we use a composite objective function that consists of a supervised component $\mathcal{L}_{\text{sup}}$ and an unsupervised component $\mathcal{L}_{\text{unsup}}$. We include the reconstruction error and the KL divergence from VAEs, and explicitly introduce an ℓ_1-norm regularization for sparsity as well as ℓ_2-norm regularization in the loss function for the unsupervised component as follows:

$$\begin{aligned} \mathcal{L}_{\text{unsup}} = \sum_{n=1}^{N} \sum_{t=1}^{T} \Big(& \mathbb{E}_{q_\phi(\mathbf{z}|\tilde{\mathbf{x}}_t^{(n)})}[\ln p_\theta(\tilde{\mathbf{x}}_t^{(n)})] \\ & -KL[q_\phi(\tilde{\mathbf{x}}_t^{(n)}|\mathbf{z})||p(\mathbf{z})] \Big) + \lambda_1|\mathbf{a}| + \lambda_2||\mathbf{a}||_2^2. \end{aligned} \tag{11}$$

where $\mathbf{a} = [\theta, \phi]^\top$ is the parameter set of the encoder and decoder, and λ_1 and λ_2 are hyperparameters.

In terms of the supervised component, we apply an asymmetric similarity loss function based on F_β scores, introduced in [14] to solve highly imbalanced dataset. The loss function l between $f(\mathbf{c}_t; \mathbf{w}, b)$ and y is defined as:

$$\begin{aligned} \mathcal{L}_{\text{sup}} &= l(f(\mathbf{c}_t; \mathbf{w}, b), y) \\ &= \frac{(1+\beta^2) \sum_{n=1}^{N} p_i g_i}{(1+\beta^2) \sum_{n=1}^{N} p_i g_i + \beta^2 \sum_{n=1}^{N} (1-p_i) g_i + \sum_{n=1}^{N} p_i (1-g_i)}. \end{aligned} \tag{12}$$

where p_i is the probability that the observed data is predicted to be positive and g_i is the ground truth label that is 1 for positive and 0 for negative. Hence, the overall loss is then defined by $\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}}$. We optimize all the parameters of our model in an end-to-end manner with stochastic gradient descent using the ADAM optimizer [4].

IV. Experiments

In this section, we validate the performance of our framework using MIMIC-III dataset. We first quantitatively evaluate the predictive performance, comparing with other methods in the literature. Furthermore, we qualitatively analyze the results of imputation using our framework.

A. Data Description

We used the publicly available MIMIC-III dataset[1] that contains longitudinal measurements for more than 40,000 critical care patients. We selected a subset of 13,998 patients with at least 48 hours of hospital stay, and sampled the time series every 2 hour in the first 48 hours. For each patient, 99 different longitudinal measurements are selected, which are divided into four main categories: laboratory measurements, inputs to patients, outputs collected from patients, and drug

[1] MIMIC-III data set is publicly available at https://mimic.physionet.org/.

TABLE I
MODEL PERFORMANCES MEASURED BY AUC AND AUPRC SCORE FOR MORTALITY PREDICTION

	Model	AUC (*mean*±*std*)	AUPRC (*mean*±*std*)
RNN Model	Zero Impute - GRU	0.7595 ± 0.0126	0.2854 ± 0.0295
	Soft Impute - GRU	0.7604 ± 0.0180	0.2340 ± 0.0282
	KNN - GRU	0.7501 ± 0.0126	0.2501 ± 0.0340
	MICE - GRU	0.7524 ± 0.0120	0.3187 ± 0.0777
	GAN - GRU	0.5189 ± 0.0169	0.1765 ± 0.1301
	AE - GRU	0.7558 ± 0.0209	0.2404 ± 0.0092
	VAE - GRU	0.7908 ± 0.0091	0.3117 ± 0.0187
	VAE+UNC - GRU	**0.7986 ± 0.0094**	**0.3201 ± 0.0272**
Non-RNN Model	Zero Impute - LR	0.7249 ± 0.0225	0.2327 ± 0.0223
	Soft Impute - LR	0.7516 ± 0.0116	0.2385 ± 0.0280
	KNN - LR	**0.7740 ± 0.0097**	0.2589 ± 0.0302
	MICE - LR	0.7356 ± 0.0103	0.2408 ± 0.0201
	GAN - LR	0.5052 ± 0.0468	0.0927 ± 0.0087
	AE - LR	0.7340 ± 0.0099	0.2257 ± 0.0138
	VAE - LR	0.6907 ± 0.0151	0.1905 ± 0.0208
	VAE+UNC - LR	0.7679 ± 0.0203	**0.2876 ± 0.0295**

prescriptions. The selected time series were scarcely observed leading to a missing rate of about 93%. For in-hospital mortality label, the ratio between positive (dead in hospital) and negative (alive in hospital) is about 1:12. We performed data cleaning by handling inconsistent units, multiple recordings at the same time, and ranged recording of feature values. We referred to [15], [17] for feature selection, data cleaning and preprocessing of MIMIC-III dataset.

B. Implementation details

All the inputs were normalized to be of zero mean and unit variance for each variable. Instead of removing abundant samples to handle data imbalance between two classes, we used bootstrapping to increase the size of the rare samples so that the number of data between the classes was balanced.

For the imputation using VAEs, the generative and recognition network were fully connected networks with 6 hidden layers, each with $\{128, 84, 72, 64, 32, 16\}$ units per layer for the encoder and the same number of units for the decoder in reverse order, using Rectified Linear Unit (ReLU) activation functions. The latent space was 16-dimensional. As for the imputation scenario of using the obtained uncertainty from the VAEs, batch normalization was further applied after the input data weighting operation using the softmax of the negative uncertainty vector. In the predictive model, we employed 1 layer GRU with 128 hidden units. The λ_1 and λ_2 were chosen from $\{0.0001, 0.0005\}$, and β from $\{1, 2\}$ based on performance on the training set. We trained our models for 250 epochs using minibatches of 150 samples with a learning rate range of {0.0001, 0.0003}. We chose the optimal parameters based on the performance for validation set.

C. Experimental Results

We report the results from the 5-fold cross validation in terms of the average Area Under the ROC curve (AUC), average Area Under Precision-Recall Curve (AUPRC) for the mortality prediction task. We compared other competing methods in the literature by grouping them into imputation methods and predictive methods. For baseline imputation methods, we included (i) Zero Impute, (ii) Soft Impute, (iii) KNN, (iv) MICE, (v) GAN, (vi) AE. The Soft Impute, KNN and MICE are implemented by using fancyimpute library[2] in python.

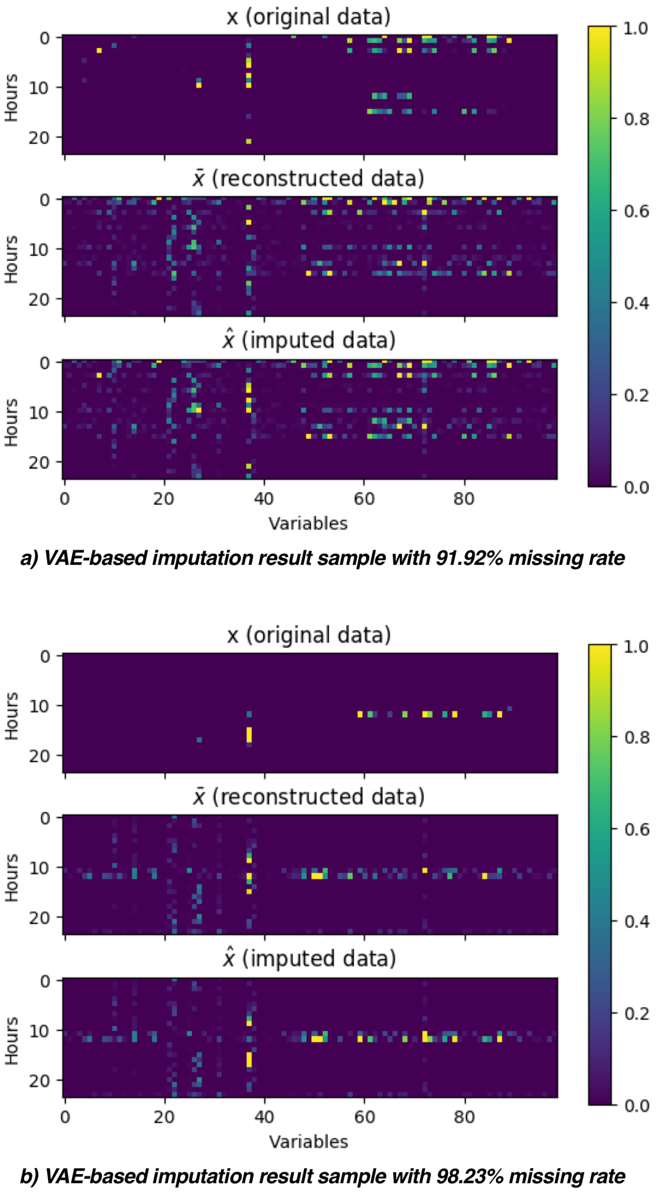
For baseline predictive methods, we categorized them into non-RNN baseline such as LR, and RNN baseline such as GRU. Since the non-RNN model cannot directly handle time series of different lengths, we regularly sample the time-series data to get a fixed length input.

Table I compares the predictive performance of the mortality depending on various imputation techniques and predictive models on MIMIC-III dataset. For the Non-RNN model, KNN imputation method using LR exhibited the highest performance in term of AUC, whereas AE and VAE without uncertainty didn't perform well, which are even lower than the simple Zero Impute. However, we could see that by utilizing the uncertainty, both AUC and AUPRC of the proposed imputation method are improved immensely. The proposed method even achieved the highest AUPRC among others. It revealed the eminent impacts of the uncertainty-aware attention mechanism to the prediction task. To confirm this outcome, it is expected to observe the imputation effects by employing the predictive model which is able to handle the time series data, since we are dealing with the high dimensional time series EHR data. In general, by employing GRU to the imputed data with various methods, we can observe the improvement of the performances. In contrast with the imputation using AE, in particular, both VAE-based imputation scenarios performed significantly higher than the others. Compared to other imputation methods, the uncertainty-aware imputation produced the highest AUC and AUPRC that indicated the accomplishment of prediction by considering the fact that the classes are imbalanced.

The learned distribution and pattern of data are essential factors in imputing the missing values. Regarding this matter, in Fig. 3, we present the results of the imputation using VAEs.

[2] The code is publicly available at https://github.com/iskandr/fancyimpute.

In order to make the comparable range of values, we set the range of values to be between zero and one. We compared the imputed samples with different missing rates. In that figures, we could observe that the VAEs are able to capture the complex pattern of the EHR data and the relations among the variables as well, even for the extreme case for data with high missing rate (98.23%) as shown in Fig. 3.b. Thus, it is proper to utilize the data for the imputation to the missing values. In addition, we also could see that the the range of the estimated imputation values follows that of the original input data in general. However, when there exist outliers in the data with extremely high (or low) values, it seems that VAEs could not achieve the same range of values.



a) VAE-based imputation result sample with 91.92% missing rate

b) VAE-based imputation result sample with 98.23% missing rate

Fig. 3. Visualization of VAE-based imputation method on different missing rates: a) 91.92% and b) 98.23%

V. Conclusions

We proposed the Bayesian framework for missing values imputation method. Based on variational inference, we exploited variational autoencoder, and further utilized the uncertainty from the generative network. By incorporating the imputed data and uncertainty, the proposed method introduced uncertainty-aware attention, which effectively utilizes the uncertainty of missing values estimation for the predictive model. In addition, we trained the overall network in an end-to-end manner, leading to effective imputation as well as optimal performances. We evaluated our method by using real-world EHR dataset, MIMIC-III, and observed higher performance than other comparative methods in mortality prediction task.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00053, A Technology Development of Artificial Intelligence Doctors for Cardiovascular Disease).

References

[1] A. Nazabal, P. M. Olmos, Z. Ghahramani and I. Valera, “Handling incomplete heterogeneous data using VAEs,” *arXiv preprint arXiv:1807.03653*, 2018.

[2] B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, “Strategies for handling missing data in electronic health record derived data,” *eGEMs*, vol. 1, no. 3, 2013.

[3] B. K. Beaulieu-Jones and J. H. Moore, “Missing data imputation in the Electronic Health Record using deeply learned autoencoders,” in *Pacific Symposium on Biocomputing*, vol. 22 pp. 207-218, 2016.

[4] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[5] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[6] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart and J. Sun, “RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism,” in *Neural Information Processing Systems (NIPS)*, pp. 3504-3512, 2016.

[7] E. Choi, A. Schuetz, W. F. Stewart and J. Sun, “Using recurrent neural network models for early detection of heart failure onset,” in *Journal of the American Medical Informatics Association (JAMIA)*, vol. 24, no. 2, pp. 361-370, 2017.

[8] J. Yoon, J. Jordon and M. Schaar, “GAIN: Missing data imputation using generative adversarial nets,” in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. 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, 2014.

[10] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple Imputation by Chained Equations: What is it and how does it work?,” in *International Journal of Methods Psychiatric Research*, vol. 20 no. 1, pp. 4049, 2011

[11] M. Lin, Q. Chen, and S. Yan, “Network in network,” In *Proceedings of the 2014 International Conference on Learning Representations (ICLR)*, 2014.

[12] M. S. Santos, J. P. Soares, P. H. Abreu1, H. Araújo, and J. Santos, “Influence of data distribution in missing data imputation,” in *Conference on Artificial Intelligence in Medicine in Europe (AIME)*, pp. 285294, 2017.

[13] S. Hochreiter and J. Schmidhuber,“Long short-term memory,” in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[14] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, “Asymmetric similarity loss function to balance precision and recall in highly unbalanced deep medical image segmentation,” *arXiv preprint arXiv:1803.11078*, 2018.

[15] S. Purushotham, C. Meng, Z. Che, and Y. Liu, “Benchmarking deep learning models on large healthcare datasets,” in *Journal of Biomedical Informatics (JBI)*, 2018.

[16] Z. C. Lipton, D. C. Kale, C. Elkan, R. Wetzel, “Learning to diagnose with LSTM recurrent neural networks, ” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[17] Z. Che, S. Purushotham, K. Cho, and Y. Liu, “Recurrent neural networks for multivariate time series with missing values in *Scientific Reports*, vol. 8, no. 6085, 2018.