2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)
August 20-24, 2022. Mexico City, Mexico

2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) | 978-1-6654-9042-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CASE49997.2022.9926629

Ensemble Generative Adversarial Imputation Network with Selective Multi-Generator (ESM-GAIN) for Missing Data Imputation

Yuxuan Li, Ayse Dogan, and Chenang Liu

Abstract— **As a pervasive issue, missing data may influence the data modeling performance and lead to more difficulties of completing the desired tasks. Many approaches have been developed for missing data imputation. Recently, by taking advantage of the emerging generative adversarial network (GAN), an effective missing data imputation approach termed generative adversarial imputation nets (GAIN) was developed. However, its modeling architecture may still lead to significant imputation bias. In addition, with the GAN structure, the training process of GAIN may be instable and the imputation variation may be high. Hence, to address these two limitations, the ensemble GAIN with selective multi-generator (ESM-GAIN) is proposed to improve the imputation accuracy and robustness. The contributions of the proposed ESM-GAIN consist of two aspects: (1) a selective multi-generation framework is proposed to identify high-quality imputations; (2) an ensemble learning framework is incorporated for GAIN imputation to improve the imputation robustness. The effectiveness of the proposed ESM-GAIN is validated by both numerical simulation and two real-world breast cancer datasets.**

Index— **Ensemble learning, GAIN, missing data imputation, multi-generator generation,**

I. Introduction

As a common data quality issue, missing data may be caused by many reasons, such as insufficient data collection and lost records. For instance, in the healthcare systems, some of the patient information may be missing, and it is also hard to revisit the patients and recover the missing information [1, 2]. More importantly, in data-driven precise disease screening and diagnosis, missing data may lead to significant bias to train the predictive models from data. Hence, it is critically needed to address the missing data issue. A common approach to address this issue is to perform data imputation [3].

In recent decades, many imputation approaches have been developed. Specifically, the missing data imputation approaches could be categorized into two groups: conventional methods and machine learning-based methods. The conventional methods include the statistics-based imputation [4], matrix completion [5], and statistical model-based approaches such as the popular expectation maximization (EM) algorithm [6]. Although they are fairly easy to calculate, the performance might be unsatisfied when the underlying data distribution is complex. Hence, the machine learning-based methods have been developed rapidly in recent years, such as the k nearest neighbors (k-NN) [7], MissForest [8], denoising autoencoder (DAE) [9], and the generative adversarial nets (GAN) [10]-based imputation approaches. However, due to the interpolation nature of k-NN, it is also not capable of handling complex data. In addition, the performance of MissForest may also be limited since it needs to run separately for each data matrix that needs to be imputed. As for DAE, it requires the complete data for training, but obtaining a complete dataset may be very challenging in real-world applications. Therefore, by taking advantage of the emerging GAN techniques, the GAN-based imputation approaches have been developed. By incorporating generator and discriminator in an adversarial learning architecture, the complex underlying data distribution could be learnt effectively without strict assumptions and complete dataset. For example, Kim *et al.* [11] has provided a detailed survey about GAN-based imputation approaches, such as the generative adversarial imputation nets (GAIN) [12], MisGAN [13], and Collaborative GAN [14]. Particularly, for the multivariate datasets, GAIN [12] is widely applied due to its superior performance than others.

In GAIN, the samples are generated and the values in the generated samples are extracted to impute the missing data. However, the values in the generated samples may be significantly different from actual values, which may lead to the imputation bias. In addition, according to the conventional GAN architecture, the training process of GAIN may also be instable and the imputation may have very high variation. Therefore, to address these gaps, a new imputation approach termed ensemble GAIN with selective multi-generator (ESM-GAIN) is proposed, and its main contributions consist of: (1) a selective multi-generation framework is proposed to identify high-quality imputations and increase imputation accuracy; and (2) an ensemble learning framework is applied for GAIN imputation to further improve the model robustness.

The rest of this paper is structured as follows. The missing data problem is defined and the GAIN is introduced in Sec. II. Then the proposed research methodology is discussed in Sec. III. Afterwards, the simulation study and a real-world case study are conducted in Sec. IV. Finally, the conclusions are discussed in Sec. V.

II. Problem Statement

Suppose that the data matrix $\mathbf{X}$ follows $\mathbb{R}^{s\times n}$. That is, $\mathbf{X}$ involves n variables and there are s samples in total. Define the mask matrix $\mathbf{M}$ in $\mathbb{R}^{s\times n}$ as a binary matrix and each element in M is shown as (1). In this way, $\mathbf{M} \odot \mathbf{X}$ represents

*This work is partially supported by the National Institutes of Health under Award Number EY 033861, and the National Science Foundation under Award Number IIP 2141184.

Yuxuan Li, Ayse Dogan, and Chenang Liu are with the School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74078 USA (*corresponding author: Chenang Liu*. e-mail: chenang.liu@okstate.edu).

the actual values in the data matrix while $(\mathbf{1}-\mathbf{M})\odot\mathbf{X}$ represents the missing values in the data matrix, where $\odot$ means the element-wise multiplication.

$$M_{ij}=\begin{cases}0 & \text{If } \mathbf{X}_{ij} \text{ is missing}\\ 1 & \text{Otherwise}\end{cases}\quad i=1,\dots,s;j=1,\dots,n \tag{1}$$

This work is based on the recently developed generative adversarial imputation network (GAIN), which has demonstrated its superior performance than the conventional imputation algorithms [12]. Based on the GAN architecture [10], GAIN also involves two components, the generator G and the discriminator D. G will generate the fake data matrix while D will distinguish whether the input values in the matrix are generated values or actual values. G and D will compete with each other. Specifically, three different matrices are sent to the GAIN model. That is, the data matrix $\mathbf{X}$, the mask matrix $\mathbf{M}$, the hint matrix $\mathbf{H}$. $\mathbf{X}$ is to record the actual values and missing values. $\mathbf{M}$ is to describe whether the values are missing or not as shown in (1). Based on the hint rate parameter h, $\mathbf{H}$ marks the area that the discriminator D should pay more attention to. In this way, based on different $\mathbf{H}$, the information passed to D will be different, which may make G learn the distributions more accurately. In each iteration, the output sent to the discriminator is shown in (2).

$$\hat{\mathbf{X}}=\mathbf{M}\odot\mathbf{X}+(\mathbf{1}-\mathbf{M})\odot G(\mathbf{X},\mathbf{M},\mathbf{Z}) \tag{2}$$

G will simulate the artificial samples, i.e., $G(\mathbf{X},\mathbf{M},\mathbf{Z})$, based on the noise Z. As shown in (2), with the help of mask matrix $\mathbf{M}$, if the element is missing, it will be imputed by the generated value. Otherwise, it will keep the original values in $\mathbf{X}$. In this way, $\hat{\mathbf{X}}$ could be obtained Then $\hat{\mathbf{X}}$ is sent to D with $\mathbf{H}$ to be distinguished. Based on the above-mentioned processes for generator and discriminator, the minimax game for GAIN is shown in (3).

$$\min_G\max_D V(D,G)=\mathbb{E}_{\hat{\mathbf{X}},\mathbf{M},\mathbf{H}}\left[\mathbf{M}^T\log\left(D(\hat{\mathbf{X}},\mathbf{H})\right)\right]+(\mathbf{1}-\mathbf{M})^T[\log(1-D(\hat{\mathbf{X}},\mathbf{H}))] \tag{3}$$

It is important to note that, only when $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ is similar to $\mathbf{M}\odot\mathbf{X}$, extracting $(\mathbf{1}-\mathbf{M})\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ as the imputed values are convincing**.** Therefore, it is essential to reduce the differences between $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ and $\mathbf{M}\odot\mathbf{X}$. In order to achieve that, the mean square error (MSE) between $G(\mathbf{X},\mathbf{M},\mathbf{Z})$ and $\mathbf{X}$, $L_M(G(\mathbf{X},\mathbf{M},\mathbf{Z}),\mathbf{X})$, is calculated as the MSE loss. Then L_M is added in the loss for G, L_G, to update the generator. Under such circumstances, L_G and the loss for D, L_D, are obtained in (4) where α is a hyper-parameter [12].

$$\begin{aligned}L_D&=(1-\mathbf{M})^T\log\left(1-D(\hat{\mathbf{X}},\mathbf{H})\right)-\mathbf{M}^T\log\left(D(\hat{\mathbf{X}},\mathbf{H})\right)\\ L_G&=-(1-\mathbf{M})^T\log\left(1-D(\hat{\mathbf{X}},\mathbf{H})\right)+\alpha L_M(G(\mathbf{X},\mathbf{M},\mathbf{Z}),\mathbf{X})\end{aligned} \tag{4}$$

Through L_M, the difference between $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ and $\mathbf{M}\odot\mathbf{X}$ could be reduced. However, after the model converges, $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ may be still significantly different from $\mathbf{M}\odot\mathbf{X}$, which means the estimation to $\mathbf{X}$, i.e., $G(\mathbf{X},\mathbf{M},\mathbf{Z})$, is still biased. Besides, due to the GAN structures [10], the training process of GAIN may be instable and the imputation variation may be high. Therefore, in order to address the two limitations of GAIN, the ensemble generative adversarial imputation network with selective multi-generator (ESM-GAIN) is proposed to implement better data imputation in Sec. III.

III. Research Methodology

In this section, the overall architecture is described in Sec. III-A. Afterwards, the selective multi-generation framework is proposed in Sec. III-B, and the ensemble learning framework for the selective multi-generator is discussed in Sec. III-C.

A. Overall architecture of the proposed ESM-GAIN

The overall architecture of the proposed ESM-GAIN is shown in Fig. 1. It involves two main novel components: (1) a new selective multi-generator framework; and (2) integrating the ensemble learning framework to GAIN. In the proposed selective multi-generator, the data matrix $\mathbf{X}$, mask matrix $\mathbf{M}$ are sent to k generators, which are applied by inputting k different random matrix $\mathbf{Z}$. Afterwards, a new selective filter layer is developed, which is partially inspired by our prior work, the augmented time-regularized GAN (ATR-GAN) [15]. All the generated samples are sent to the selective filter layer to identify the generated samples which are similar to the actual samples. Then both the selected samples and the hint matrix $\mathbf{H}$ obtained from $\mathbf{M}$ are sent to the discriminator D. In this way, the losses for both generators and discriminators could be estimated, and the selective multi-generator could be updated as well.

As shown in Fig. 1, each trained GAIN with a selective multi-generator could output one imputed matrix. Afterwards, the ensemble learning framework is then incorporated. If the element to be imputed is continuous, the mean of such element from all the imputed matrices is used. Otherwise, the median is used. In this way, all the missing values could be imputed.

B. Selective multi-generation framework

B.1. Selective filter layer

As described in Sec. II, the missing values in the actual samples are replaced by the values from the same location in the artificial samples. However, $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ may be different from $\mathbf{M}\odot\mathbf{X}$. For each sample, the variables may be correlated. Hence, any minor changes from the actual values may lead to large differences between imputed values. Under such circumstances, when $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ is not the same as $\mathbf{M}\odot\mathbf{X}$, it is not convincing to impute $(\mathbf{1}-\mathbf{M})\odot\mathbf{X}$ by $(\mathbf{1}-\mathbf{M})\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$. Since data matrix $\mathbf{X}$ follows $\mathbb{R}^{s\times n}$, denote that the generator may also generate s samples with n variables as $\overline{\mathbf{X}}$. In order to make $\mathbf{M}\odot G(\mathbf{X},\mathbf{M},\mathbf{Z})$ closer to $\mathbf{M}\odot\mathbf{X}$, the selective filter layer, L, is proposed in this work.

Definition 1. **(Selective filter layer)**: Selective filter layer L is designed to select the artificial samples from $\overline{\mathbf{X}}$ that is similar as $\mathbf{X}$. Based on the one-to-one Euclidean distance calculation between the samples in $\overline{\mathbf{X}}$ and $\mathbf{X}$, L can select the desired samples by judging whether the distance is less than the threshold δ, formulated with an indicator function I

$$L=\overline{\mathbf{X}}\cdot I_{\{d(\overline{\mathbf{X}},\mathbf{X})<\delta\}}(\overline{\mathbf{X}}) \tag{5}$$

Based on the selective filter layer, the input of D, $\hat{\mathbf{X}}^*$, could be obtained as shown in (6). The generated samples that are more similar to actual data will be selected as $\overline{\mathbf{X}}^*$. Then the corresponding actual data, $\mathbf{X}^*$, will be combined with $\overline{\mathbf{X}}^*$ as $\hat{\mathbf{X}}^*$,

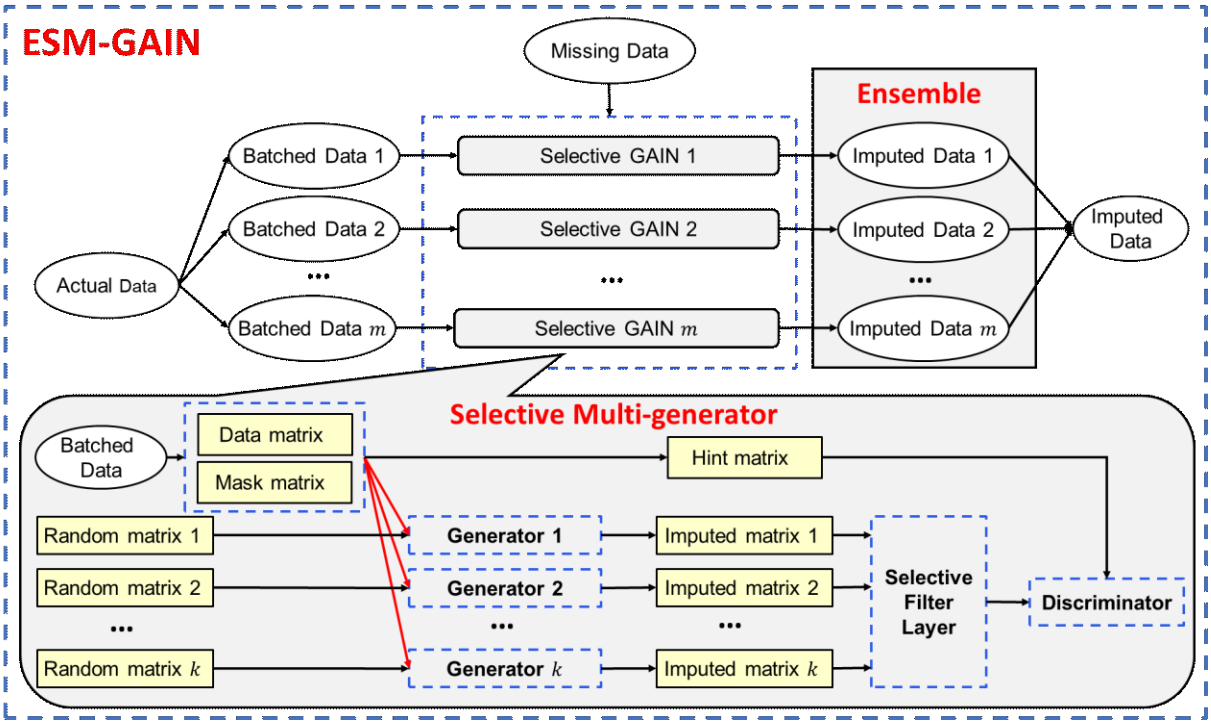

Figure 1 A demonstration of the proposed ESM-GAIN framework.

based on the adjusted mask matrix, $\mathbf{M}^*$. Afterwards, $\hat{\mathbf{X}}^*$ could be sent to the discriminator to update the entire model.

$$\begin{aligned} \bar{\mathbf{X}}^* &= L(G(\mathbf{X}, \mathbf{M}, \mathbf{Z})|\delta) \\ \hat{\mathbf{X}}^* &= \mathbf{M}^* \odot \mathbf{X}^* + (1 - \mathbf{M}^*) \odot \bar{\mathbf{X}}^* \end{aligned} \tag{6}$$

B.2. Multi-generator collaboration

Based on the selective filter layer, high-quality samples could be obtained. However, due to the existence of δ, the number of samples in $\bar{\mathbf{X}}^*$ is much less than the number of samples in $\bar{\mathbf{X}}$. Since the actual data should have the same samples as $\bar{\mathbf{X}}^*$, it is not guaranteed that all the actual samples in $\mathbf{X}$ could be selected. Besides, due to the sample size deduction, the diversity of samples sent to the discriminator is also limited. Hence, in order to increase the number of actual samples which could be selected and improve the diversity of imputed samples, a multi-generator is applied.

Denote that there are k generators, $\{G_1, G_2, \dots, G_k\}$, to generate artificial samples. Thus, in each iteration, k groups of imputed samples, $\{\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \dots, \bar{\mathbf{X}}_k\}$, will be generated. Since the selective filter layer does not have any neural network parameters, one common selective filter layer could be applied for all the generators simultaneously with the input actual samples, $\mathbf{X}$. Hence, based on the selection layer, $\{\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \dots, \bar{\mathbf{X}}_k\}$ are transformed to $\{\bar{\mathbf{X}}_1^*, \bar{\mathbf{X}}_2^*, \dots, \bar{\mathbf{X}}_k^*\}$. Then with the corresponding actual samples $\{\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_k^*\}$, the samples that combine the actual values and imputed values, $\{\hat{\mathbf{X}}_1^*, \hat{\mathbf{X}}_2^*, \dots, \hat{\mathbf{X}}_k^*\}$, could be obtained based on (7).

$$\hat{\mathbf{X}}_i^* = \mathbf{M}_i^* \odot \mathbf{X}_i^* + (1 - \mathbf{M}_i^*) \odot \bar{\mathbf{X}}_i^* \quad i = 1,2, \dots, k \tag{7}$$

$\{\hat{\mathbf{X}}_1^*, \hat{\mathbf{X}}_2^*, \dots, \hat{\mathbf{X}}_k^*\}$ could be sent to the discriminator to distinguish whether the values are actual values or imputed values. In this way, the output of the discriminator could be applied to calculate the losses for all D and $\{G_1, G_2, \dots, G_k\}$. As shown in (8). the discriminator needs to consider the average output of all the generators in L_D. However, since the generated samples from one generator are not related to other generators, each generator has its own loss as shown in (8).

$$\begin{aligned} L_D = \sum_{i=1}^{k} \frac{1}{k} ((1 - \mathbf{M}_i^*)^T \log\left(1 - D(\hat{\mathbf{X}}_i, \mathbf{H}_i)\right) - \\ \mathbf{M}_i^{*T} \log\left(D(\hat{\mathbf{X}}_i, \mathbf{H}_i)\right)) \\ L_{G_i} = -(1 - \mathbf{M}_i^*)^T \log\left(1 - D(\hat{\mathbf{X}}_i, \mathbf{H}_i)\right) + \\ \alpha L_M(G_i(\mathbf{X}, \mathbf{M}, \mathbf{Z}), \mathbf{X}) \;\; i = 1,2, \dots, k \end{aligned} \tag{8}$$

When the losses converge, the model is considered as well-trained and could be applied for data imputation. Afterwards, the k imputed data matrices from the generators, $\{\hat{\mathbf{X}}_1^*, \hat{\mathbf{X}}_2^*, \dots, \hat{\mathbf{X}}_k^*\}$, are obtained. Therefore, the imputed data matrix of the model, $\hat{\mathbf{X}}$, could be obtained by calculating the mean of each element in the imputed matrices as shown in (9). Specifically, rounding will be applied to transform the calculated values of discrete elements to integers.

$$\hat{\mathbf{X}} = \sum_{i=1}^{k} \frac{1}{k} \hat{\mathbf{X}}_i^* \tag{9}$$

There are two hyperparameters to be determined in the proposed selective multi-generator, i.e., δ and k. Specifically, to control the sample size for the output of the selective filter layer, δ could be obtained based on the percentile of calculated distance [15]. In this way, the number of samples sent to the discriminator will be the same for each iteration. Afterwards, δ and k could be determined based on cross validation [15]. Under different δ and k, the model is trained and the mean absolute error (MAE) of the imputation is calculated. Then the δ and k with the smallest MAE is selected for the imputation.

C. Incorporation of ensemble learning

By integrating the proposed selective filter layer and multi-generator generation, the proposed selective multi-generator is expected to improve the accuracy of imputation. To further improve model robustness, an ensemble learning framework is further incorporated in the proposed method, i.e., ESM-GAIN.

In the ensemble learning framework, m data matrices, $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$, will be obtained from the data matrix $\mathbf{X}$ through bootstrapping. m is a hyperparameter that could be determined by cross validation. Based on the data matrices, m GAINs with selective multi-generator are trained separately. In this way, the ensemble learning framework could learn the actual distribution more comprehensively. In addition, during the training process, each GAIN with selective multi-generator may apply different h to increase the diversity among different models. After the training, $\mathbf{X}$ is sent to each model and the imputed matrices, $\{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \dots, \hat{\mathbf{X}}_m\}$, could be obtained.

Notably, besides the values of the elements, the labels of the elements, i.e., continuous or discrete, are also sent to G since the continuous and discrete elements to be imputed will be considered separately. If the element is continuous, the mean of such elements from different imputed data matrices is calculated. On the other hand, if the element is discrete, the median of such elements from different imputed data matrices is selected. By calculating the mean/median, the model robustness could be improved since the inappropriate imputed values, i.e., outliers, will not interfere with the final output. In this way, the final imputed data matrix, i.e., $\mathbf{X}'$, could be obtained and output from the ensemble learning framework.

The overall algorithm for ESM-GAIN is shown below. Based on the bootstrapped data, the GAINs with selective multi-generator are trained. Afterwards, the entire data matrix is sent to each selective multi-generator framework to obtain the imputed data matrices. Finally, the imputed data matrices are combined to obtain and output the final imputed matrix.

Algorithm 1: ESM-GAIN algorithm

Input: Actual data matrix $\mathbf{X}$, Parameter m, k, s and δ
Step 1: Bootstrap data matrix $\mathbf{X}$ to $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$
For $i = 1$ **to** m **do**
  For $j = 1$ **to** k **do**
    Step 2: Randomly choose s actual samples $\mathbf{X}_i^j$ from actual sample set $\mathbf{X}_i$
    Step 3: Generate s artificial samples $\bar{\mathbf{X}}_i^j$ from generator G
    Step 4: Send $\bar{\mathbf{X}}_i^j$ to the selection layer L to obtain $\bar{\mathbf{X}}_i^{j*}$
    Step 5: Obtain $\hat{\mathbf{X}}_i^{j*}$ based on $\bar{\mathbf{X}}_i^{j*}$ and $\mathbf{X}_i^j$
  Step 6: Send $\hat{\mathbf{X}}_i^{1*}, \hat{\mathbf{X}}_i^{2*}, \dots, \hat{\mathbf{X}}_i^{k*}$ into discriminator D to get output $D(\hat{\mathbf{X}}_i^{1*})$, $D(\hat{\mathbf{X}}_i^{2*})$, …, $D(\hat{\mathbf{X}}_i^{k*})$
  Step 7: Optimize the model parameters based on the output of discriminator
  Until L_{G_1}, L_{G_1}, …, L_{G_k} and L_D converge:
    Step 8: Send $\mathbf{X}$ to $\{G_1, G_2, \dots, G_k\}$ and impute as $\hat{\mathbf{X}}$
Step 9: Get mean/median from $\{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \dots, \hat{\mathbf{X}}_m\}$ as $\mathbf{X}'$
Output: $\mathbf{X}'$

IV. Case Studies

To validate the effectiveness of the proposed ESM-GAIN, both numerical simulation (Sec. IV-A) and a real-world case study in healthcare (Sec. IV-B), are applied. The effectiveness of the proposed data imputation approach can be represented by the quality of the imputed values. Thus, in this study, the mean absolute errors (MAE) after data imputation are applied as an evaluation metric to validate the effectiveness of the proposed ESM-GAIN in missing data imputation.

A. Simulation study

In the simulation study, the Gaussian process (GP) is applied to simulate 2000 actual samples with 30 variables, based on the radial basis function (RBF) kernel. Two values, 0.001 and 0.005, are applied for the parameter θ in the RBF kernel. Besides, to make the simulation data closer to real-world cases, noises are also added. The process to simulate the data are shown in (10).

$$\begin{gathered}\mathbf{X} = \mathbf{2} \times \begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix} + \begin{pmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{pmatrix}, \mathbf{X}_i = \begin{bmatrix}\mathbf{x}_1\\ \cdots\\ \mathbf{x}_{1000}\end{bmatrix}, \mathbf{Z}_i = \begin{bmatrix}\mathbf{z}_1\\ \cdots\\ \mathbf{z}_{1000}\end{bmatrix},\\ \mathbf{x}_j \sim GP(0, \kappa), z_{jl} \sim N(0, 2^2),\\ \kappa(x_{jl_1}, x_{jl_2}) = \exp\left(-\frac{1}{2\theta_i}(\|x_{jl_1} - x_{jl_2}\|_2^2\right)\\ \theta_1 = 0.001, \theta_2 = 0.005,\\ i = 1,2, j = 1,2, \dots,1000, l, l_1, l_2 = 1,2, \dots,30\end{gathered} \tag{10}$$

In this way, the 2000×30 data matrix is generated. In order to demonstrate the effectiveness of the proposed method under discrete variables, each variable in the data matrix is categorized into five levels from 1 to 5. The setups of parameters are shown in Table I.

TABLE I. The Data and Parameter Setups

Setup	Value
Sample size	2000×30
Number of generators k	2
Number of selective multi-generators m	10
Threshold δ	80th percentile of the calculated distance
Batch size s	128

The neural network structures of ESM-GAIN follow Yoon, *et al.* [12]. Since the variables turn to discrete variables, the median is calculated as the output in the ensemble learning framework. Besides, some benchmark methods, including MissForest [8], matrix completion [5], the k-NN imputation algorithm [7], and GAIN [12], are applied to better test the performance of the proposed method. The experiment for each approach is conducted three times and the average MAE is calculated. To make the validation more comprehensively, the missing value rate increases from 20% to 50% while the sample size increases from 800 to 2000.

Before comparing the proposed ESM-GAIN with other data imputation methods, it is important to first test the effectiveness of the proposed selective multi-generator and the incorporated ensemble learning framework. Hence, the ensemble GAIN (no selective multi-generator) and GAIN are also applied as ablation experiments.

The MAEs under different missing value rates and sample size are shown in Fig. 2. The MAEs of ensemble GAIN are mostly smaller than the MAEs of GAIN, showing that the

ensemble learning framework is effective. The MAEs of ESM-GAIN are much smaller than the other two methods under each missing value rate. Hence, the newly added components are effective to improve the imputation accuracy. Besides, as the missing value rates increase, the MAEs for all three approaches increase. Such a pattern is normal since higher missing value rates means less information for imputation.

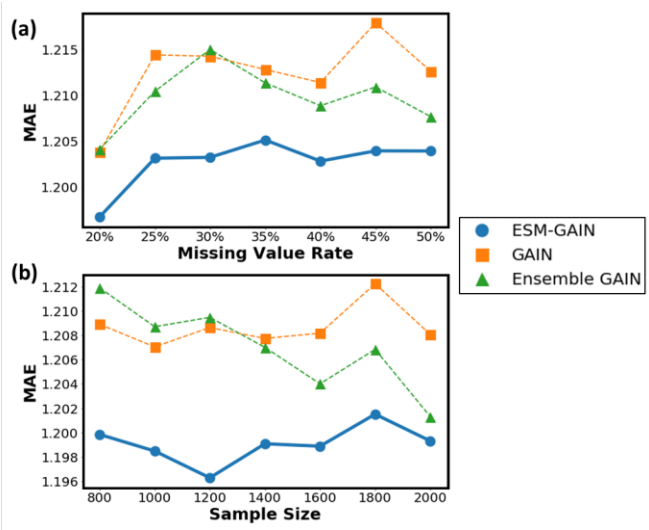

Figure 2 MAE comparisons between ESM-GAIN, ensemble GAIN and GAIN under different missing value rates (a) and sample size (b).

As shown in Fig. 2(b), compared with GAIN, the MAEs of ensemble GAIN are smaller when the sample size is larger than 1200. However, the MAEs of ensemble GAIN are still comparable since there are no significant differences between ensemble GAIN and GAIN when the sample size is less than 1200. Besides, the MAEs of ESM-GAIN are always the smallest under each sample size. Hence, the effectiveness of newly added components is validated.

In addition, the proposed method is also compared with the benchmark methods under different missing value rates and sample size. Notably, the setup of missing value rates and sample size are the same as the setup for validating the components in ESM-GAIN. The MAEs for different approaches are shown in Fig. 3.

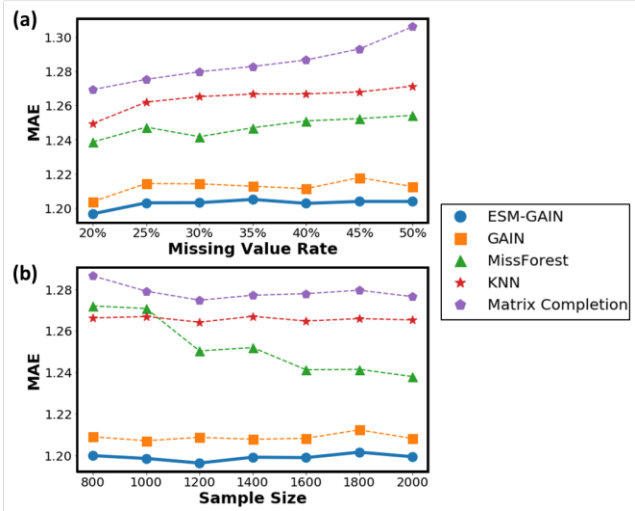

Figure 3 MAE comparisons between ESM-GAIN, GAIN, MissForest, *k*-NN, matrix completion under different missing value rates (a) and sample size (b).

As shown in Fig. 3(a), the MAEs of *k*-NN imputation, matrix completion and MissForest are much higher than the MAEs of GAIN and ESM-GAIN. Besides, compared with the MAEs of GAIN, the MAEs of ESM-GAIN are also lower under each missing value rate. In addition, as the missing value rate increases, the MAEs of all the benchmark methods also increase. However, the MAEs of the proposed method do not significantly increase, which also shows the imputation stability of the proposed method.

Furthermore, as shown in Fig. 3(b), the MAEs of *k*-NN imputation, matrix completion and MissForest are still much higher than GAIN and ESM-GAIN. Besides, the MAEs of the ESM-GAIN are also lower than GAIN. Hence, it demonstrates that the proposed method always has the smallest MAE under each sample size and the proposed method could also have satisfied stable performance. Overall, the simulation study demonstrates the outperformance of the proposed method.

B. Real-world case study

In this section, in order to validate the effectiveness of the proposed method for both continuous and discrete elements, two datasets for the breast cancer from the UCI database [16] are applied: (1) wisconsin diagnostic breast cancer (WDBC) dataset [17]; and (2) the breast-cancer-wisconsin dataset [18].

The WDBC dataset involves 569 samples and each sample involves 30 continuous variables. Each sample records the computed features from one image of the cell nucleus, including radius, texture, perimeter, and so on. The detailed descriptions about the variables could be found in [17].

Besides, the breast-cancer-wisconsin dataset involves 699 samples and each sample involves 9 discrete variables. Each sample also records the features of one cell, including clump thickness, cell size, cell shape and so on [18]. However, different from the WDBC dataset, each discrete variable involves 10 levels, i.e., from 1 to 10.

Yoon *et al.* [12] has validated that GAIN can perform much better than the common imputation methods, including MissForest and matrix completion for the breast cancer data. Thus, this case study mainly focuses on the comparisons between the proposed method and GAIN. For the WDBC data set, since the data contains continuous variables, the mean values for each element that need to be imputed are calculated in the ensemble learning framework. As for the breast-cancer-wisconsin dataset, the median values for each element that needs to be imputed are calculated. The other setups of the proposed method in this case are the same as Sec. IV-A. The average and standard deviations of MAEs for the ESM-GAIN and GAIN are shown in Table II as the missing value rate is 20%.

TABLE II. THE IMPUTATION PERFORMANCE OF ESM-GAIN AND GAIN IN TERMS OF MAE (AVERAGE ± STD OF MAE)

Algorithms	WDBC dataset	Breast-cancer-wisconsin dataset
GAIN	0.0526 ± 0.0010	1.1700 ± 0.0289
ESM-GAIN	$\mathbf{0.0505 \pm 0.0006}$	$\mathbf{1.1027 \pm 0.0123}$

As shown in Table II, for both datasets, the average MAEs of the proposed method are smaller than the average MAEs of

the GAIN. In addition, the MAE standard deviations of ESM-GAIN are also smaller under both datasets, which shows that the proposed method is more stable than GAIN. Hence, it shows that the proposed method is more effective than GAIN for both continuous and discrete datasets.

To further demonstrate the effectiveness of the proposed method, the comparisons when missing value rate increases from 20% to 50% are conducted as well. The MAEs of GAIN and the proposed method are shown in Fig. 4. For the WDBC dataset, as shown in Fig. 4(a), the MAEs of ESM-GAIN are always less than the MAEs of GAIN. Specifically, as the missing value rate increases, the MAEs are also increasing smoothly. Such a pattern also proves that higher missing value rates lead to less information for data imputation, resulting in higher MAEs.

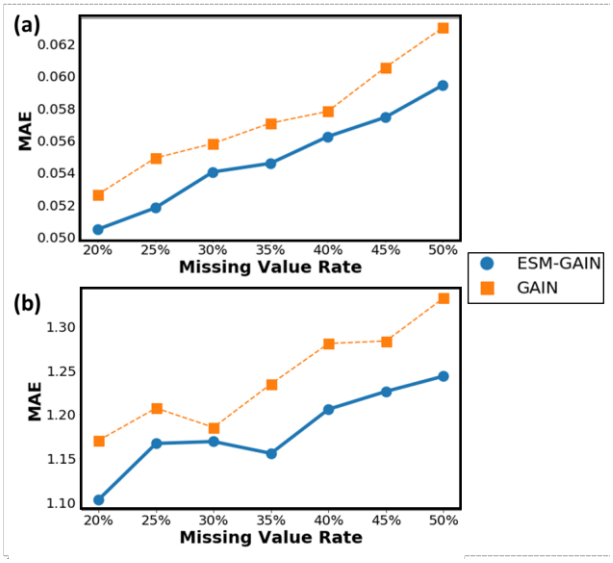

Figure 4 The MAE comparisons between ESM-GAIN, and GAIN under different missing value rates based on the WDBC dataset (a) and breast-cancer-wisconsin dataset (b) from the breast cancer data.

Besides, for the breast-cancer-wisconsin dataset, as shown in Fig. 4(b), the MAEs of ESM-GAIN are also smaller than the MAEs of GAIN. Though it is not as smooth as Fig. 4(a), it is still able to show the trend that the MAEs may increase as the missing value rate increases. Overall, the real-world case study also demonstrates that the proposed method outperforms GAIN for both continuous and discrete datasets.

V. Conclusion

In this paper, a new data imputation approach termed ensemble GAIN with selective multi-generator (ESM-GAIN) is proposed to impute the missing values in the datasets. Compared with GAIN, it involves two significant advantages: (1) a selective multi-generator framework is proposed to identify the generated samples and improve the imputation accuracy; (2) the ensemble learning framework is applied to the GAINs with selective multi-generator to further improve the model robustness.

In this study, the superior performance of ESM-GAIN over the benchmark methods is demonstrated by both numerical simulation data and the breast cancer datasets from UCI data. In addition, to validate the effectiveness of newly added components, the performance of GAIN structure excluding one or two components is compared against the ESM-GAIN. Under different sample sizes and missing value rates, the imputation MAEs of the proposed method can always be reduced effectively. Besides, the real-world case study also shows the outperformance of the proposed method. Thus, the proposed method is very promising for both continuous and discrete data imputation.

References

[1] Yoon, J., Davtyan, C., & van der Schaar, M. (2016). Discovery and clinical decision support for personalized healthcare. *IEEE journal of biomedical and health informatics, 21*(4), 1133-1145.

[2] Dogan, A., Li, Y., Odo, C. P., Sonawane, K., Lin, Y., & Liu, C. (2022). A Utility-Based Machine Learning-Driven Personalized Lifestyle Recommendation for Cardiovascular Disease Prevention. *medRxiv*.

[3] Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy, 18*(2), 2308-2316.

[4] Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research, 24*(7), 815-829.

[5] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, *11*, 2287-2322.

[6] García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, *19*(2), 263-282.

[7] Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), 2541-2552.

[8] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

[9] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103).

[10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[11] Kim, J., Tae, D., & Seok, J. (2020, February). A survey of missing data imputation using generative adversarial networks. In *2020 International Conference on Artificial Intelligence in Information and Communication* (ICAIIC) (pp. 454-456). IEEE.

[12] Yoon, J., Jordon, J., & Schaar, M. (2018, July). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning* (pp. 5689-5698). PMLR.

[13] Li, S. C. X., Jiang, B., & Marlin, B. (2018, September). MisGAN: Learning from Incomplete Data with Generative Adversarial Networks. In *International Conference on Learning Representations*.

[14] Lee, D., Kim, J., Moon, W. J., & Ye, J. C. (2019). CollaGAN: Collaborative GAN for missing image data imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2487-2496).

[15] Li, Y., Shi, Z., Liu, C., Tian, W., Kong, Z., & Williams, C. B. (2021). Augmented Time Regularized Generative Adversarial Network (ATR-GAN) for Data Augmentation in Online Process Anomaly Detection. *IEEE Transactions on Automation Science and Engineering.*

[16] Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[17] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570-577.

[18] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", *Proceedings of the National Academy of Sciences, U.S.A., Volume 87*, December 1990, pp 9193-9196.