Applied Intelligence (2024) 54:11068–11082
https://doi.org/10.1007/s10489-024-05814-2

Improved generative adversarial imputation networks for missing data

Xiwen Qin[1] · Hongyu Shi[1] · Xiaogang Dong[1] · Siqi Zhang[1] · Liping Yuan[1]

Accepted: 24 August 2024 / Published online: 5 September 2024

Abstract

Conventional statistical methods for missing data imputation have been challenging to adapt to the large-scale new features of high dimensionality. Moreover, the missing data imputation methods based on Generative Adversarial Networks (GAN) are plagued with gradient vanishing and mode collapse. To address these problems, we have proposed a new imputation method based on GAN to enhance the accuracy of missing data imputation in this study. We refer to our missing data method using Generative Adversarial Imputation Networks (MGAIN). Specifically, the least squares loss is first introduced to solve the gradient vanishing problem and ensure the high quality of the output data in MGAIN. To mitigate mode collapse, dual discriminator is used in the model, which improved the diversity of output data to avoid the degradation of computational performance caused by single data. As a result, MGAIN generates rich and accurate imputation values. The MGAIN enhances imputation accuracy and reduces the root mean square error metric by 21.66% compared to the baseline model. We evaluated our method on baseline datasets and found that MGAIN outperformed state-of-the-art and popular imputation methods, demonstrating its effectiveness and superiority.

Keywords Missing Data Imputation · Generative Adversarial Networks · Least squares · Dual discriminator

1 Introduction

The issue of missing data is a significant problem in statistical analysis across all statistical applications. Various reasons for missing data include mechanical and human reasons. Data play a crucial role in artificial intelligence, and high-quality data directly affect the quality of knowledge output. Since the quality of data is an essential indicator of their value, a substantial amount of data in a real production setting can lead to quality problems. Moreover, a large amount of low-quality data can lower the density of values in the data, potentially resulting in a biased final analysis that does not fully leverage information inherent in the data. Therefore, correctly and efficiently handling missing data is critical [1, 2].

The current state-of-the-art missing data imputation algorithms fall into two main categories: discriminative algorithms and generative algorithms. Discriminative algorithms include Multiple Imputation by Chained Equations (MICE) [3], MissForest [4], and matrix completion [5]. Furthermore, generative algorithms include K-NearestNeighbor (KNN) [6], Expectation Maximization (EM) [7], and machine learning-based algorithms such as Auto-encoders (AE) [8] and GAN [9]. However, current generative algorithms have some limitations. They are based on data distribution assumptions [10], making them less suitable for datasets containing mixed categories and continuous variables.

In previous statistical studies, the sample sizes were often small and the proportion of missing data was low. Researchers usually use subjective judgment to discard or manually process the missing records. However, with the advent of the big data era, the proportion of missing data has become larger because data dimensions have skyrocketed. Manual processing is inefficient at this point. Discarding missing records leads to a loss of a significant amount of

✉ Xiaogang Dong
dongxiaogang@ccut.edu.cn

Xiwen Qin
qinxiwen@ccut.edu.cn

Hongyu Shi
2227678478@qq.com

Siqi Zhang
1634728618@qq.com

Liping Yuan
2218269062@qq.com

[1] School of Mathematics and Statistics, Changchun University of Technology, 130012 Yan'an Street, Changchun, Jilin, China

information [11], leading to systematic differences between incomplete and complete observations. Analyzing such data is likely to lead to wrong conclusions. As data volume and dimension increase, statistical learning methods are widely used in daily classification [12, 13], prediction [14, 15], and dimensionality reduction [16]. However, when the original input data of statistical learning methods contain missing values, most statistical learning methods will be unusable. Only specialized methods such as methods based on decision trees (Cox Regression Model Combined with Decision Tree [17], Branch-Exclusive Splits Trees (BEST) [18], and Other variants of decision trees [19]) and methods based on random forests (Depth-Weighted Prevalence for an random forests Tree Ensemble [20], PhyloMissForest [21], Logistic Ridge Regression and Random Forest Ensemble Model [22], and Random forest with the assignation of missing entries [23]) can handle missing values, but they still have limitations, such as significant degradation in the prediction accuracy.

Imputation of missing data has always been a difficult and hot problem. Various computational methods have been proposed to address missing data in multiple fields, including medical [24], image mapping [25], and financial data [26, 27]. In a study tackling the missing MRI problem, Sharma et al. used the GAN and designed a multi-input, multi-output network model to fill in the missing information. GAN was found to have good applicability in multimodal problems [28].

Using machine learning methods to handle missing data has become an active research. However, existing machine learning methods for processing missing data still face limitations. A new approach called Generative Adversarial Imputation Networks (GAIN) has gained attention since its proposal in 2018 [29]. GAIN leverages the capabilities of GAN to learn the original data distributions for imputation, surpassing the traditional statistical methods and other machine learning methods in dealing with missing data. Awan [30] proposed a new method for computing missing data based on class-specific features to address the challenges of modeling a single distribution over the entire dataset. However, this method ignores the problem of class-specific features of the data. Conditional Generative Adversarial Imputation Networks (CGAIN) address this issue using class-specific distributions for missing data, producing the best estimate of missing values.

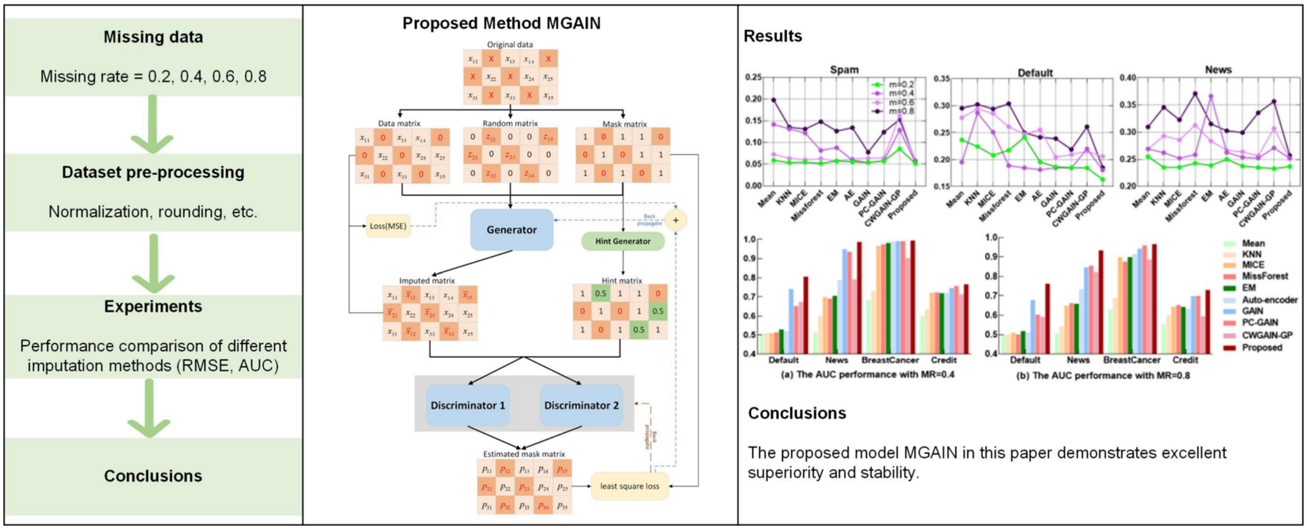
Additionally, PC-GAIN, a new unsupervised missing data computation method, addresses the problem of GAIN, overlooking the issue of the latent class information reflecting the relationship between samples. PC-GAIN uses the latent class information to further improve the accuracy of filling in the missing values [31]. The Generative Adversarial Guider Imputation Network proposed in 2022 focuses on unsupervised interpolation to handle the locally homogeneous regions, particularly at the boundaries [32]. Wu et al. proposed a method based on the Fuzzy c-Means algorithm and GAIN to exploit the information on the local samples [33]. Zhao et al. introduced an imputation method called Multiple Generative Adversarial Imputation Networks based on data attributes [34]. A study also uses deep metric learning and MisGAN methods for multi-tasking missing data imputation [35]. For time series continuous missing values, Wang et al. proposed Wasserstein GAN with gradient penalty (CWGAIN-GP) [36]. However, the discriminator of WGAN-GP usually fails to maintain continuity in the peripheral region of the true sample distribution.

However, the above GAIN-based data imputation methods address the original issues of the original gradient vanishing and mode collapse in GAN. This limitation can affect the model performance and lead to model overfitting. Moreover, the data that are filled out may lack high quality and diversity. For example, the generators of PC-GAIN and MisGAN may get stuck in a dead-end loop of generating similar samples, resulting to a lack of diversity in filling results. CWGAIN-GP is sensitive to noise or outliers, leading to unstable or inaccurate filling results. This study proposes MGAIN to overcome these challenges. Unlike traditional imputation methods and other popular imputation methods, MGAIN incorporates least squares loss to address the gradient vanishing and uses dual discriminator to mitigate mode collapse, ensuring high quality and diversity of the output data. Therefore, the motivation of this paper is to utilize the MGAIN to address the problems of traditional imputation methods and GAN-based approaches, enhancing the quality and diversity of missing data imputation, and offering a more reliable and effective solution for real data processing tasks.

Missing data are classified mainly into three categories, including completely random missing, random missing, and nonrandom missing [37]. Completely random missing data refer to data that are entirely missing at random, independent of incomplete or complete variables, and do not introduce bias to the sample. Random missing data are the probability of missing data not related to the missing data but only to the partially observed data. The nonrandom missing data occur when the missingness is related to the values taken by the incomplete variables. This paper focuses on missing completely at random (MCAR) data.

Therefore, we followed the approach depicted in Fig. 1. First, we processed the original datasets with four missing rates of 0.2, 0.4, 0.6, and 0.8. Subsequently, the missing data was subjected to data preprocessing. Then, the data were compared using different imputation methods. Finally, conclusions were drawn based on the experiments and results. The main contributions of this paper are as follows,

1. A novel missing data imputation method, MGAIN, has been developed to address the issue of missing data. This



Missing data

Missing rate = 0.2, 0.4, 0.6, 0.8

Dataset pre-processing

Normalization, rounding, etc.

Experiments

Performance comparison of different imputation methods (RMSE, AUC)

Conclusions

Proposed Method MGAIN

Results

Conclusions

The proposed model MGAIN in this paper demonstrates excellent superiority and stability.

Fig. 1 Graphical abstract

approach address the challenges associated with incomplete data.

2. A combination of least squares loss function and dual discriminator has been proposed to overcome the challenges of mode collapse and gradient vanishing in GAN. This innovative approach effectively solves these problems, and theoretically demonstrates its feasibility.
3. We extensively tested the proposed model on diverse datasets to assess its effectiveness and generalizability. These experimental results demonstrate the model's potential to address the challenges of missing data estimation in various practical applications.

The remainder of this paper is organized as follows. In Section 2, we review GAN and its variants. Section 3 provides a detailed overview of the proposed method. Section 4 presents the theoretical analysis. Section 5 presents the experimental methodology and the experimental results. Finally, Section 6 summarizes the paper.

2 Methodology

In this section, we mainly review the model principles and computational problems based on GAN, which mainly focus on our research work. This section mainly introduces GAN, Least Squares Generative Adversarial Networks (LSGAN) and Dual Discriminator Generative Adversarial Networks (D2GAN) as shown in Fig. 2. LSGAN and D2GAN are variants of GAN. We utilize variants of GAN for missing data imputation. To the best of our knowledge, this is the first time that both variants have been applied to missing data.

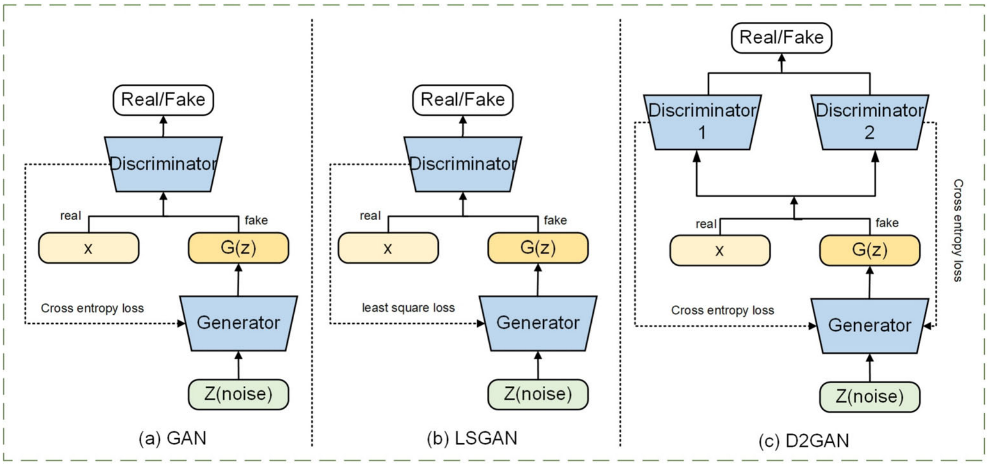


Fig. 2 Framework diagram of GAN and its variants

2.1 GAN

Goodfellow (2014) introduced a GAN consisting of a generator and a discriminator (Fig. 2(a)), both of which are composed of multilayer perceptrons [9]. The generator learns the original data distribution to create artificial samples (called fake samples) that are similar to real samples, while the discriminator distinguishes the difference between real and generated samples. The core logic of all GAN networks is that generator and discriminator play against each other. And generation is mainly embodied in the two-player games until the generator can generate the discriminator can not determine the real or fake samples.

$x \sim P_{\text{data}}$ is the original data distribution, and $z \sim p_z$is generally a noise that obeys a uniform or normal distribution. The generator and discriminator are abbreviated as G and D, respectively. z is used as input to G, which is then fed into an output $G(z)$, which is then fed into a D. Then the input to D is two parts x and $G(z)$, and tries to determine whether the input is real or fake. Therefore, the discriminator is a binary classification problem using a Sigmoid function that produces outputs in the range between 0 and 1. The GAN framework corresponds to an extremely large and extremely small two-player game with the following objective function $L(D, G)$:

$$\min_G \max_D L(D, G) = E_{x \sim P_{data}}[\log D(x)] + E_{z \sim P_z}[\log(1 - D(G(z)))]. \tag{1}$$

$$L_G = E_{z \sim P_z}[\log(1 - D(G(z)))]. \tag{2}$$

$$L_D = -E_{x \sim P_{data}}[\log D(x)] - E_{z \sim P_z}[\log(1 - D(G(z)))]. \tag{3}$$

where L_D and L_G are the objective functions D of G and , respectively.

As a powerful deep learning model, GAN has achieved great success in many tasks, but it also has some shortcomings and challenges. During training, the generator may fall into mode collapse, resulting in a lack of diversity in the generated samples and too many repetitions of the same patterns. This can limit the quality of the generated results. Similarly, gradient vanishing is a common problem that hinders the convergence and training effectiveness of the model and tends to lead to training instability. Overall, GAN, while making significant progress in generative tasks, still face challenges and require further research and improvements to address these issues.

2.2 LSGAN

The LSGAN is a variant of GAN proposed in 2017 [38]. It is shown in Fig. 2(b). The cross-entropy loss function used in the original GAN leads to the problem of gradient vanishing, and to solve this problem, LSGAN takes the least squares loss function network. And LSGAN has two benefits over conventional GAN. First, LSGAN can generate higher-quality images. Second, LSGAN is more stable in the learning process. The following is the objective function of LSGAN:

$$\min_D L_D = \frac{1}{2} E_{x \sim P_{x \sim data}} \left[(D(x) - b)^2\right] + \frac{1}{2} E_{z \sim P_z} \left[(D(G(z)) - a)^2\right]. \tag{4}$$

$$\min_G L_G = \frac{1}{2} E_{z \sim P_z} \left[D(G(z) - c)^2\right]. \tag{5}$$

where L_D and L_G are the objective functions of D and G, respectively. a and b are the labels of the generated and real data, respectively. c denotes the value of the generated data that the generator wants the discriminator to believe. In this paper, $a = 0$ and $b = c = 1$.

Compared with traditional GAN, LSGAN is more stable in the training process, which avoids the common training instability problem in traditional GAN. LSGAN has a positive impact on both the generator and the discriminator, which improves the overall training effect and the quality of the generated results. However, LSGAN still suffers from mode collapse, resulting in a lack of diversity in the generated results. On the whole, LSGAN has some advantages over traditional GAN in terms of stability and quality of generated results. However, LSGAN also has some disadvantages, which need to be weighed and selected according to the specific tasks and datasets. For some specific generation tasks, LSGAN may be an valid choice.

2.3 D2GAN

D2GAN [39], proposed by Nguyen et al. (2017), differs from the original GAN in that D2GAN contains not only one generator but also two discriminators. As shown in (c) in Fig. 2. It combines the Kullback-Leibler (KL) divergence and reverse KL divergence into a unified objective function, thus utilizing the complementary statistical properties of these divergences to effectively disperse the estimation density and alleviate the mode collapse problem. The objective function $L\left(G, D_1, D_2\right)$ of D2GAN can be expressed as follows:

$$\begin{aligned}\min_G \max_{D_1, D_2} L\left(G, D_1, D_2\right) =& \alpha \times E_{x \sim P_{data}}\left[\log D_1(x)\right] + E_{z \sim P_z}\left[-D_1(G(z))\right] \\ &+ E_{x \sim P_{data}}\left[-D_2(x)\right] + \beta \times E_{z \sim P_z}\left[\log D_2(G(z))\right]\end{aligned}. \tag{6}$$

where D_1, D_2 and G denote discriminator 1, discriminator 2, and generator, respectively. α and β are hyperparameters, $0 < \alpha \leq 1, 0 < \beta \leq 1$. The role of α and β is to control the effect of KL divergence and inverse KL divergence on the optimization problem.

D2GAN can efficiently learn multimodal data distributions through the triple confrontation of the generator and the two discriminators. D_1 and D_2 are giving high scores to the distribution P_{data} from the original data and P_G from the generated data and vice versa. D_1 and D_2 do not share their parameters.

By introducing two discriminators, D2GAN can evaluate the realism of the generated samples more efficiently, thus facilitating the generation of more realistic and high-quality samples by the generator. Moreover, D2GAN can also reduce the risk of mode collapse and avoid the generator from falling into local optimal solutions for generating repeated samples. However, D2GAN still faces some difficulties, such as gradient vanishing and training instability. Overall, D2GAN improves the performance of the generator, which brings higher-quality generated samples, but it also needs to face challenges such as increased training complexity.

The current GAN and its various variants are riddled with problems such as gradient vanishing, modal collapse, and training instability, limiting their performance and applications. We propose a new GAN variant to address these issues, improving model stability and generation. We apply this new GAN variant to dealing with missing data. Applying MGAIN to missing data can effectively improve data generation and increase the model's ability to handle missing data. This innovative GAN variant introduces new ideas and methods to address missing and incomplete data in practical problems and has a wide range of application prospects and research value.

3 Proposed method

In this section, a noteworthy GAIN is proposed, which is a new method to deal with missing data based on GAN. Yoon et al take each value in incomplete data whether it is missing or not, as a category label constitutes the missing mask, and combined with Conditional GAN (CGAN), they propose GAIN model to realize the imputation of missing data, and experimentally illustrate that the method is better than the traditional imputation method. However, GAIN suffers from the common problems of GAN, i.e., gradient vanishing and mode collapse problems. Therefore, to solve the above problems, a new imputation method based on GAN is proposed in this paper. As shown in Fig. 3. This structure and theory are described in detail below.

3.1 Inputs of the model

First, define the original data as $X = (X_1, X_2, X_3, ..., X_n)$. X is a random variable in the n-dimensional space $\mathbb{X} = (\mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3 \times ... \times \mathbb{X}_n)$, and n represents the total number of samples. Define M to be the mask vector, $M = (M_1, M_2, M_3, ..., M_n)$, which takes values from $\{0, 1\}^n$. Define a new random variable $\tilde{X} = \left(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, ..., \tilde{X}_n\right)$, $\tilde{X}$ is a random variable in the n-dimensional space $\tilde{\mathbb{X}} = \left(\tilde{\mathbb{X}}_1 \times \tilde{\mathbb{X}}_2 \times \tilde{\mathbb{X}}_3 \times ... \times \tilde{\mathbb{X}}_n\right)$, which is obtained from the following equation:

$$\tilde{X}_i = \begin{cases} X_i, if M_i = 1 \\ *, otherwise \end{cases} . \tag{7}$$

where $M_i = 1$ means that these data are not missing, otherwise it is missing.

The purpose of the calculation method is to calculate the missing values in $\tilde{X}$. The purpose of the missing data calculation is to generate samples based on the conditional probability of X given $\tilde{X}$, i.e., $P\left(X|\tilde{X} = \tilde{X}_i\right)$.

3.2 Generator

From Fig. 2 we can see that the input to the generator consists of the trio of data matrix X, mask matrix M, and random matrix Z(aka noise). Similar to the data and mask matrices, the random matrix is a n-dimensional vector $Z = (Z_1, Z_2, Z_3, ..., Z_n)$.

The generator G is equivalent to a mapping function of $\tilde{\mathbb{X}} \times \{0, 1\} \times [0, 1]^n \to \mathbb{X}$. Now define two random variables:

$$\begin{aligned} \bar{X} &= G\left(\tilde{X}, M, (1 - M) \odot Z\right) \\ \hat{X} &= M \odot \tilde{X} + (1 - M) \odot \bar{X}. \end{aligned} \tag{8}$$

where $\odot$ denotes the element level multiplication, $\bar{X}$ denotes the missing value portion computed by the generator, and $\hat{X}$ is composed of two variables, X and $\bar{X}$. That is, the missing value portion is filled by the $\bar{X}$ inferred by the generator, and the rest of the unmissing data consist of the original observations.

3.3 Discriminator

Unlike other GAN-based applications for missing data, the method proposed in this paper consists of two discriminators. For example, GAIN has only one discriminator; however, there are two generators (generator, hint generator). Then the mode collapse problem arises, so in order to mitigate the mode collapse problem, the dual discriminator approach is used in this study.

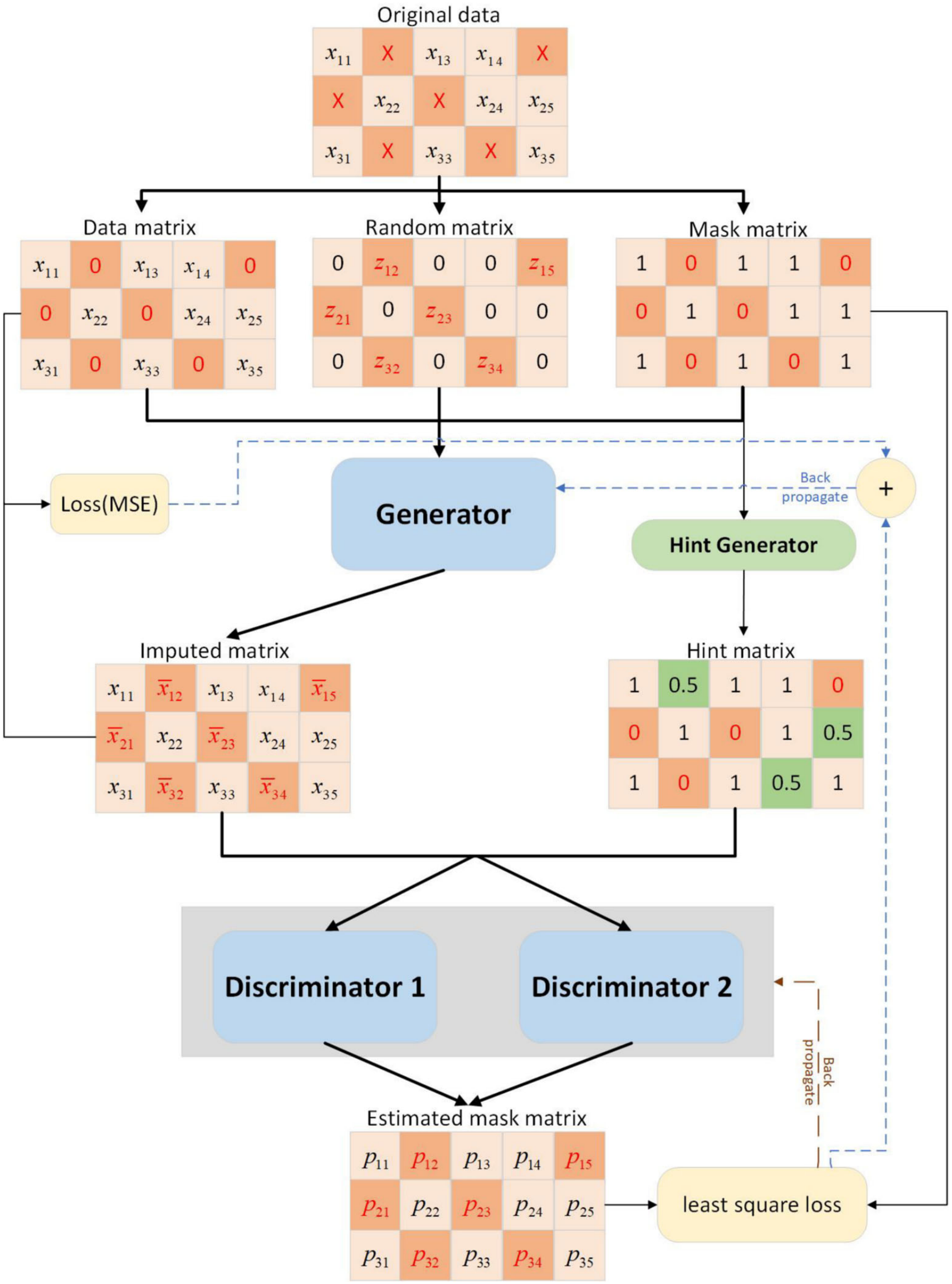

Original data
x_{11} X x_{13} x_{14} X
X x_{22} X x_{24} x_{25}
x_{31} X x_{33} X x_{35}
Data matrix
x_{11} 0 x_{13} x_{14} 0
0 x_{22} 0 x_{24} x_{25}
x_{31} 0 x_{33} 0 x_{35}
Random matrix
0 z_{12} 0 0 z_{15}
z_{21} 0 z_{23} 0 0
0 z_{32} 0 z_{34} 0
Mask matrix
1 0 1 1 0
0 1 0 1 1
1 0 1 0 1
Loss(MSE)
Generator
Back propagate
+
Hint Generator
Imputed matrix
x_{11} $\overline{x}_{12}$ x_{13} x_{14} $\overline{x}_{15}$
$\overline{x}_{21}$ x_{22} $\overline{x}_{23}$ x_{24} x_{25}
x_{31} $\overline{x}_{32}$ x_{33} $\overline{x}_{34}$ x_{35}
Hint matrix
1 0.5 1 1 0
0 1 0 1 0.5
1 0 1 0.5 1
Discriminator 1
Discriminator 2
Back propagate
Estimated mask matrix
p_{11} p_{12} p_{13} p_{14} p_{15}
p_{21} p_{22} p_{23} p_{24} p_{25}
p_{31} p_{32} p_{33} p_{34} p_{35}
least square loss

Fig. 3 Framework of MGAIN

The discriminator of the GAN in dealing with missing data does not determine whether the whole variable is true (the output is 0 or 1), but rather determines which of the components are observations and which are generated. In this process, then, it is equivalent to de-predicting the mask vector. Similarly to the generator, the discriminator D is equivalent to being a mapping function of $\mathbb{X} \to [0, 1]^n$. The ith component of $D\left(\hat{x}\right)$ is the probability that the ith element of $\hat{x}$ is observed.

3.4 Hint Generator

Similarly to the hint mechanism of Yoon et al., our approach includes a hint mechanism. The cue is represented as a random variable H which takes values in the cue space $\mathbb{H}$. The cue vector supports D by telling D some input and observation values, allowing D to decide whether to input other values or observe other values. Similar to G and D earlier, H is a mapping function of $\mathbb{X} \times \mathbb{H} \to [0, 1]^n$. The hinting mechanism is necessary because G can produce multiple distributions, and for all of them, D cannot distinguish between real and false values. Thus, giving H to D restricts the solution to a single distribution. H is obtained using (9).

$$H = B \odot M + 0.5 \odot (1 - B)\,. \tag{9}$$

where $B \in \{0, 1\}^n$ is the random variable obtained by uniformly sampling k from $\{0, 1, 2, ..., n\}^n$ and applying (10).

$$B_i = \begin{cases} 1, if\ i \neq k \\ 0, if\ i = k \end{cases}\,. \tag{10}$$

3.5 Objective function

Inspired by GAN and its variants as well as GAIN, the objective function of our MGAIN method has two parts. Second, we train G to minimize the probability that D predicts m correctly. The overall objective function, and loss function of the MGAIN method are given in (11) and (12).

$$\min_{G}\max_{D_1, D_2} V\left(G, D_1, D_2\right). \tag{11}$$

$$\begin{aligned} V\left(G, D_1, D_2\right) =& \frac{\alpha}{2} \times E_{x\sim P_{data}}\left[M\left(D_1(\hat{X}, H) - 1\right)^2\right] \\ &+ \frac{1}{2} E_{z\sim P_z}\left[(1 - M)\left(D_1(\hat{X}, H)\right)^2\right] \\ &+ \frac{1}{2} \times E_{x\sim P_{data}}\left[M\left(D_2(\hat{X}, H) - 1\right)^2\right]. \\ &+ \frac{\beta}{2} \times E_{z\sim P_z}\left[(1 - M)\left(D_2(\hat{X}, H)\right)^2\right] \end{aligned} \tag{12}$$

where, α, β are hyperparameters, $0 < \alpha, \beta \leq 1$. The role of α and β is to control the effect of minimizing the loss on the optimization problem. D_1 and D_2 represent discriminator 1 and discriminator 2, respectively. D_1 and D_2 are the ones that give high scores to the data from the original data distribution P_{data} and to the data from the generated data distribution P_G, and vice versa. Where D_1 and D_2 do not share their parameters.

D2GAN can mitigate mode collapse, but cannot avoid the problem of vanishing gradient and instability. LSGAN can solve the problem of gradient vanishing, but it is difficult to avoid the problem of mode collapse, which is the lack of diversity in the generated samples. Therefore, this paper draws on the advantages and disadvantages of these two and combines D2GAN and LSGAN to propose a new generative adversarial inference network, a model that not only mitigates mode collapse, but also solves the gradient vanishing problem.

The same as [29] the loss function of G consists of two parts, the loss of estimates and the loss of observations. Unlike it, the loss function of G consists of two discriminators and the least squares loss. The combined loss function V_G is given in (13).

$$\begin{aligned} V_G =& \frac{1}{2} E_{z\sim P_z}\left[(1 - M)\left(D_1(\hat{X}, H)\right)^2\right] \\ &+ \frac{\beta}{2} E_{z\sim P_z}\left[M\left(D_2(\hat{X}, H)\right)^2\right]. \\ &+ \lambda \sum_{i=1}^{n} m_i L_{obs}(x_i, x_i'). \end{aligned} \tag{13}$$

where λ and β are the hyperparameters, m_i is the element in the mask matrix M. In this paper, $\lambda = 0.4$. $L_{obs}(x_i, x_i')$ is defined by (14):

$$L_{obs}(x_i, x_i') = \begin{cases} (x_i, x_i')^2, if\ x_i\ is\ continuous \\ -x_i log(x_i'), if\ x_i\ is\ binary \end{cases}\,. \tag{14}$$

To be concrete, the pseudo-code of the algorithm of MGAIN is shown in Algorithm 1.

Algorithm 1 MGAIN

```
Input: missing dataset $T$, $D_1$ batch size $n_{D_1}$, $D_2$ batch size $n_{D_2}$, $G$ batch size $n_G$
Output: Final complete data and Trained algorithm
while training loss does not converge do
    1. Discriminator1 and Discriminator2 optimization
    Draw $T_D$ samples from the dataset $\{(\tilde{x}(i), m(i))\}_{i=1}^{T_D}$
    Draw $T_D$ i.i.d samples, $\{z(i)\}_{i=1}^{T_D}$ of $Z$
    Draw $T_D$ i.i.d samples, $\{b(i)\}_{i=1}^{T_D}$ of $B$
    for $i = 1, 2, ..., T_D$ do
        $\bar{x}(i) \leftarrow G(\tilde{x}(i), m(i), z(i))$
        $\hat{x}(i) \leftarrow m(i) \odot \tilde{x}(i) + (1 - m(i)) \odot \bar{x}(i)$
        $h(i) \leftarrow b(i) \odot m(i) + 0.5(1 - b(i))$
    end for
    Update $D_1$, $D_2$ using stochastic gradient descent (SGD)
    2. Generator optimization
    Draw $T_G$ samples from the dataset $\{(\tilde{x}(i), m(i))\}_{i=1}^{T_G}$
    Draw $T_G$ i.i.d samples, $\{z(i)\}_{i=1}^{T_G}$ of $Z$
    Draw $T_G$ i.i.d samples, $\{b(i)\}_{i=1}^{T_G}$ of $B$
    for $i = 1, 2, ..., T_G$ do
        $h(i) \leftarrow b(i) \odot m(i) + 0.5(1 - b(i))$
    end for
    Update $G$ using SGD with fixed $D_1$ and $D_2$
end while
```

4 Theoretical analysis

Now provide a formal theoretical analysis of our proposed model, essentially showing that given G, D_1 and D_2 have sufficient capacity, i.e., in the nonparametric limit, at the optimal point, G can recover the data distribution by minimizing the divergence between the model and the data distribution. Consider first the optimization problem of a (w.r.t.) discriminator given a fixed generator.

Proposition 1 *Minimizing $V(G, D_1, D_2)$ for a given generator yields the following closed-form optimal discriminator:*

$$D_1^* = \frac{\alpha M P_{data}}{\alpha M P_{data} + (1-M)P_z}$$
$$D_2^* = \frac{M P_{data}}{M P_{data} + \beta(1-M)P_z}$$

where, $V(G, D_1, D_2)$ is the overall objective function of MGAIN. D_1^* and D_2^* are the optimal value of D_1 and D_2. α and β are hyperparameters, same as for α and β as mentioned in Section 3.5 M is the mask matrix mentioned in Section 3.2 P_{data} and P_z represent both the raw data distribution and the noise distribution, respectively.

Proof

$$\begin{aligned} V(G, D_1, D_2) =& \frac{\alpha}{2} E_{x\sim P_{data}}\left[M\left(D_1(\hat{X}, H)-1\right)^2\right] \\ &+\frac{1}{2} E_{z\sim P_z}\left[(1-M)\left(D_1(\hat{X}, H)\right)^2\right] \\ &+\frac{1}{2} E_{x\sim P_{data}}\left[M\left(D_2(\hat{X}, H)-1\right)^2\right] \\ &+\frac{\beta}{2} E_{z\sim P_z}\left[(1-M)\left(D_2(\hat{X}, H)\right)^2\right] \end{aligned}$$

$$\begin{aligned} &=\int \frac{\alpha}{2} P_{data}(x)\left[M\left(D_1(\hat{X}, H)-1\right)^2\right]dx \\ &+\int \frac{1}{2} P_z(z)\left[(1-M)\left(D_1(\hat{X}, H)\right)^2\right]dz \\ &+\int \frac{1}{2} P_{data}(x)\left[M\left(D_2(\hat{X}, H)-1\right)^2\right]dx \\ &+\int \frac{\beta}{2} P_z(z)\left[(1-M)\left(D_2(\hat{X}, H)\right)^2\right]dz \end{aligned}$$

Let $\frac{\partial V(G,D_1,D_2)}{\partial D_1} = 0$, $\frac{\partial V(G,D_1,D_2)}{\partial D_2} = 0$, the optimal D_1 and D_2, that is, D_1^* and D_2^*, are obtained.

In the following, fixing $D_1 = D_1^*$ and $D_2 = D_2^*$, and then going to compute $V\left(G, D_1^*, D_2^*\right)$ yields the optimal G^* for generator G. □

5 Experiments

5.1 Datasets and evaluation metrics

5.1.1 Datasets

We tested the proposed MGAIN method using several publicly available real-world datasets provided by the University of California, Irvine (UCI) Machine Learning Repository and a database of handwritten digits provided by Yann. These datasets are listed in Table 1. We compare our method with the state-of-the-art GAIN method and other popular imputation methods. We also evaluated it on different proportions of missing data, ranging from 20% to 80%. In all experiments, the missing data were randomly deleted as MCAR.

The missing rate (MR) is the missing rate of the data, which can be expressed by the following equation:

$$MR = \frac{Number\ of\ missing\ values}{Total\ number\ of\ samples}. \tag{15}$$

Table 1 Datasets

Datasets	Instances	Source	Categorial variables	Numerical variables
Spam	4601	UCI	1	57
Letter	20000	UCI	0	16
MNIST	60000	Github	10	0
Default	30000	UCI	10	14
News	39644	UCI	23	35
Breast Cancer	569	UCI	9	10
Credit	30000	UCI	9	14
Air quality	6941	UCI	0	12
Wine quality	4898	UCI	1	11
Beijing Air quality	503	UCI	0	6

Table 2 Main parameters of MGAIN

Main parameters	parameter values
Batch size	128
epochs	10000
Hint rate	0.9
MR	0.2, 0.4, 0.6, 0.8
Activation function (Output Layer)	Sigmoid
Number of Convolutional Layers	3
λ	0.4
α	0.5
β	1

5.1.2 Evaluation metrics

The experimental results are based on the use of real-world datasets. We use the root mean square error (RMSE) to evaluate the experimental computational performance results:

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y^i - \hat{y}^i)^2}. \tag{16}$$

where, y^i and $\hat{y}^i$ are real values and generated values respectively. N is the number of data. To evaluate the models, this work compares the RMSE of all models on test data. This is the same evaluation metrics used by Yoon et al. [29] and Stefenon et al. [40].

Experimental prediction performance results are evaluated using Area Under the Receiver Operating Characteristic Curve (AUC). AUC is defined as the area under the ROC curve. where, the ROC curve is called the receiver operating characteristic curve, with the TPR as the vertical coordinate and the FPR as the horizontal coordinate. TPR is the true positive rate, i.e., the proportion of true positive cases that are correctly predicted to be positive cases; FPR is the false positive rate, i.e., the proportion of true negative examples that are incorrectly predicted to be positive.

$$TPR = \frac{TP}{TP + FN}. \tag{17}$$

$$FPR = \frac{FP}{FP + TN}. \tag{18}$$

Table 3 RMSE for differentand α and β

		Spam	Letter	Default	News	Breast Cancer	Credit	MNIST	Air Quality	Wine Quality	Beijing Air Quality
$\alpha = 0.1$	$\beta = 0.2$	0.0551	0.1247	0.2002	0.2440	0.0897	0.1975	0.1046	0.0802	0.1164	0.1263
	$\beta = 0.4$	0.0571	0.1261	0.1889	0.2541	0.0911	0.1947	0.1113	0.0833	0.1279	0.1148
	$\beta = 0.6$	0.0533	0.1393	0.2023	0.2477	0.0852	0.2065	0.1108	0.0889	0.1237	0.1037
	$\beta = 0.8$	0.0546	0.1281	0.2009	0.2423	0.0960	0.2023	0.1102	0.0864	0.1432	0.1010
	$\beta = 1$	0.0533	0.1273	0.2061	0.2473	0.0749	0.2041	0.1112	0.0912	0.1211	0.0764
$\alpha = 0.3$	$\beta = 0.2$	0.0549	0.1290	0.2005	0.2460	0.0933	0.1954	0.1003	0.1032	0.1123	0.1046
	$\beta = 0.4$	0.0549	0.1251	0.2029	0.2526	0.0903	0.2033	0.1081	0.0916	0.1131	0.0749
	$\beta = 0.6$	0.0551	0.1309	0.1872	0.2417	0.0831	0.2002	0.1135	0.0857	0.1044	0.1377
	$\beta = 0.8$	0.0568	0.1308	0.2000	0.2697	0.0915	0.1956	0.1131	0.1125	0.1065	0.1991
	$\beta = 1$	0.0517	0.1243	0.2052	0.2525	0.0871	0.2033	0.1110	0.0913	0.1110	0.0828
$\alpha = 0.5$	$\beta = 0.2$	0.0566	0.1257	0.1947	0.2582	0.0842	0.2011	0.0998	0.0848	0.1170	0.0752
	$\beta = 0.4$	0.0581	0.1274	0.1925	0.2448	0.0806	0.2008	0.0997	0.0814	0.1216	0.0765
	$\beta = 0.6$	0.0531	0.1285	0.1891	0.2465	0.0748	0.1979	0.0999	0.0923	0.1078	0.0808
	$\beta = 0.8$	0.0566	0.1271	0.2054	0.2519	0.0871	0.1965	0.0990	0.0867	0.1146	0.0749
	$\beta = 1$	**0.0517**	**0.1214**	**0.1633**	**0.2366**	**0.0709**	**0.1956**	**0.0969**	**0.0759**	**0.1036**	**0.0725**
$\alpha = 0.7$	$\beta = 0.2$	0.0563	0.1291	0.1955	0.2403	0.0868	0.2006	0.0996	0.0841	0.1200	0.0888
	$\beta = 0.4$	0.0568	0.1256	0.1958	0.2448	0.0994	0.1982	0.0980	0.0802	0.1115	0.0779
	$\beta = 0.6$	0.0521	0.1241	0.2034	0.2454	0.0712	0.2029	0.0980	0.0889	0.1306	0.1279
	$\beta = 0.8$	0.0548	0.1252	0.1884	0.2426	0.0924	0.1995	0.0996	0.0873	0.1117	0.1084
	$\beta = 1$	0.0543	0.1250	0.1961	0.2413	0.0826	0.1976	0.0989	0.0834	0.1103	0.0906
$\alpha = 0.9$	$\beta = 0.2$	0.0535	0.1273	0.2049	0.2518	0.0754	0.2026	0.1110	0.0831	0.1264	0.1005
	$\beta = 0.4$	0.0533	0.1253	0.2043	0.2564	0.0927	0.2052	0.1101	0.1011	0.1139	0.1027
	$\beta = 0.6$	0.0540	0.1242	0.1932	0.2492	0.0765	0.2075	0.1081	0.1098	0.1121	0.0953
	$\beta = 0.8$	0.0535	0.1619	0.2063	0.2526	0.0858	0.2052	0.1066	0.0940	0.1182	0.0789
	$\beta = 1$	0.0544	0.1305	0.2062	0.2442	0.0861	0.2038	0.1146	0.0831	0.1249	0.1860

Table 4 RMSE for MR=0.2

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Spam	0.0589	0.0537	0.0546	0.0518	0.0575	0.0556	0.0540	0.0577	0.0854	**0.0517**
Letter	0.1845	0.1931	0.1900	0.1948	0.1927	0.1356	0.1317	0.1414	0.1351	**0.1214**
Default	0.2364	0.2242	0.2079	0.2172	0.2413	0.1956	0.1864	0.1851	0.1846	**0.1633**
News	0.2549	0.2351	0.2352	0.2428	0.2386	0.2499	0.2375	0.2348	0.2327	**0.2306**
Breast Cancer	0.1174	0.1316	0.1320	0.1359	0.1360	0.2098	0.1008	0.1243	0.1560	**0.0709**
Credit	0.1977	0.1998	0.2051	0.2047	0.2103	0.1959	0.2006	0.1998	0.2008	**0.1956**
MNIST	0.1795	0.1669	0.2082	0.2234	0.1629	0.1449	0.1523	0.1562	0.1664	**0.0969**
Air Quality	0.1519	0.1631	0.1606	0.1637	0.1634	0.0989	0.1146	0.1061	0.1143	**0.0759**
Wine Quality	0.1618	0.1479	0.1499	0.1549	0.1500	0.1679	0.1094	0.1107	0.1534	**0.1036**
Beijing Air Quality	0.1026	0.1906	0.1950	0.1863	0.1873	0.2734	0.0841	0.0836	0.0946	**0.0725**

where, TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative in the samples, respectively. Higher AUC values correspond to higher model classification accuracy.

5.2 Design of experiments

In order to compare and investigate with existing missing value imputation methods, 10 datasets and 8 existing sampling methods were selected for comparison to validate the effectiveness of MGAIN. The number of missing regions is determined based on a specified missing rate. Subsequently, these regions are randomly assigned as missing regions based on the defined missing rate to generate complete data. We performed a 5-fold cross-validation in our experiments. Each experiment is repeated ten times, and its average performance is reported.

In Table 2, Batch size is the batch size, epochs is the number of iterations of the network, the hint rate is the percentage of the hinting mechanism, and MR is the missing rate of (15). α and β are the hyperparameters in (12). λ is the hyperparameters in (13). Moreover, during the training process of the MGAIN model, the Adam optimizer was used in this paper.

5.3 Quantitative analysis of MGAIN

First, we performed experiments to find the appropriate α and β then compared RMSE and AUC with other missing data imputation methods. This experiment was conducted when the missing rate was 0.2, and the specific experimental results are shown in Table 3. The best results among all the experimental results in this section are highlighted in bold.

According to Table 3 we can observe that the proposed model in this paper has the best performance at $\alpha = 0.5, \beta = 1$. Therefore, $\alpha = 0.5, \beta = 1$ is chosen for the following comparison experiments with other methods.

We compared our proposed MGAIN method with Mean, KNN, MICE, MissForest, EM, Auto-encoder, GAIN, PC-GAIN, and CWGAIN-GP methods. Mean, KNN, MICE and EM are popular statistical imputation methods, whereas, in a recent study, GAIN, MissForest, Auto-encoder, PC-GAIN, and CWGAIN-GP completions were the best-performing

Table 5 RMSE for MR=0.4

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Spam	0.1418	0.1313	0.1219	0.0818	0.0883	0.0594	0.0539	0.0579	0.1287	**0.0538**
Letter	0.2600	0.2666	0.2685	0.2746	0.2712	0.1428	0.1359	0.1354	0.1840	**0.1328**
Default	0.1953	0.2870	0.2507	0.1887	0.1848	0.1813	0.1846	0.1845	0.2199	**0.1808**
News	0.2691	0.2621	0.2520	0.2585	0.3666	0.2626	0.2537	0.2522	0.1891	**0.0964**
Breast Cancer	0.1663	0.1814	0.1888	0.1936	0.1923	0.2216	0.1054	0.1111	0.1560	**0.0709**
Credit	0.2007	0.1968	0.2007	0.1994	0.1968	0.2005	0.2045	0.1980	0.2083	**0.1952**
MNIST	0.1996	0.2013	0.2319	0.2557	0.1991	0.1903	0.2063	0.1765	0.1937	**0.1236**
Air Quality	0.1717	0.2217	0.2290	0.2311	0.2288	0.1908	0.1250	0.1242	0.1372	**0.1057**
Wine Quality	0.1866	0.2070	0.2109	0.2152	0.2147	0.1781	0.1514	0.1243	0.1989	**0.1084**
Beijing Air Quality	0.1379	0.2519	0.2625	0.2709	0.2666	0.3165	0.0983	0.0940	0.1294	**0.0912**

Table 6 RMSE for MR=0.6

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Spam	0.0728	0.0634	0.0604	0.0626	0.0579	0.0619	0.0639	0.0642	0.1625	**0.0555**
Letter	0.3189	0.3310	0.3286	0.3352	0.3266	0.1665	0.1459	0.1588	0.1995	**0.1414**
Default	0.2778	0.2946	0.2854	0.2609	0.2474	0.2552	0.2046	0.2094	0.2155	**0.2060**
News	0.2689	0.2929	0.2857	0.3129	0.2836	0.2666	0.2637	0.2565	0.3068	**0.2516**
Breast Cancer	0.2019	0.2122	0.2296	0.2373	0.2222	0.2333	0.1060	0.1466	0.2192	**0.1059**
Credit	0.2087	0.2133	0.2132	0.2237	0.2099	0.2214	0.2247	0.1899	0.2251	**0.1896**
MNIST	0.2116	0.2351	0.2558	0.2551	0.2097	0.2118	0.2176	0.1887	0.1975	**0.1484**
Air Quality	0.1837	0.2703	0.2801	0.2779	0.2734	0.2508	0.1510	0.1384	0.1774	**0.1272**
Wine Quality	0.2070	0.2555	0.2568	0.2607	0.2542	0.1895	0.1636	0.1507	0.2101	**0.1272**
Beijing Air Quality	0.1732	0.3134	0.3231	0.3240	0.3212	0.3160	0.1498	0.1277	0.1455	**0.1089**

methods. We performed experiments at missing rates of 0.2, 0.4, 0.6, and 0.8, and the results are shown in Tables 4, 5, 6 and 7.

The proposed MGAIN model had exhibited the best performance with MR=0.2. For example, MGAIN reduced the RMSE by up to 0.2029 on the Beijing Air Quality dataset, because EM did not effectively handle the missing values when faced with time series datasets. CWGAIN-GP was introduced as the method to deal with missing time series data in 2024, indicating that MGAIN was better than CWGAIN-GP, could handle tabular data, and had some generalization ability on time series data.

Table 5 presents the results of RMSE at MR=0.4. The proposed model outperformed classical and GAIN-based methods. For example, our models demonstrated superiority on the Spam and BreastCancer datasets, reducing 0.088 and 0.1158, respectively, compared with the worst model. Additionally, we found that the errors of all models were generally larger on the two datasets with the largest amount of data (MNIST and News) than on the other datasets. However, our models performed best on these two datasets, demonstrating the capability of MGAIN to handle large amount of data.

At MR=0.6, the experimental results demonstrated the effectiveness of the proposed method in dealing with high missing rate data. First, the RMSE of MGAIN was smaller than that of all other methods. For example, on the News dataset, the RMSE of MGAIN is 0.0049 lower than that of the second-best model, PC-GAIN. Second, the missing rate ranged from 0.2 to 0.6, with a smaller increase in MGAIN. For instance, with MR=0.2, MGAIN's RMSE on the Letter dataset increaseed by 0.02 from 0.1214 to 0.1414, while MissForest's RMSE on the Letter dataset increases by 0.2404 from 0.1948 to 0.3352, suggesting that MGAIN can handle missing data with stability.

We also conducted experiments with MR=0.8, and the specific results are shown in Table 7. MGAIN performed excellently when MR=0.8. MGAIN decreased by 0.1776 compared with the latest model CWGAIN-GP on the Beijing Air Quality dataset. This result indicates that CWGAIN-GP is unsuitable for handling data with a high missing rate. During the experiments, the PC-GAIN running time was longer than that of other models. Therefore, the three Spam, Default, and News datasets are visualized as examples to better view the proposed model's superiority, as shown in Fig. 4.

Table 7 RMSE for MR=0.8

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Spam	0.1980	0.1356	0.1309	0.1481	0.1264	0.1341	0.0775	0.1238	0.1532	**0.0571**
Letter	0.3682	0.3805	0.3770	0.3869	0.3707	0.1890	0.2065	0.2510	0.2530	**0.1703**
Default	0.2952	0.3025	0.2940	0.3038	0.2499	0.2415	0.2389	0.2188	0.2609	**0.1853**
News	0.3096	0.3458	0.3227	0.3712	0.3452	0.3026	0.2992	0.3357	0.3567	**0.2569**
Breast Cancer	0.2298	0.2399	0.2582	0.2685	0.2425	0.2540	0.2253	0.2832	0.3317	**0.1635**
Credit	0.2229	0.2248	0.2129	0.2182	0.2127	0.2218	0.2312	0.2156	0.2479	**0.2081**
MNIST	0.2443	0.2607	0.2179	0.2691	0.2309	0.2501	0.3024	0.2139	0.3016	**0.2035**
Air Quality	0.2429	0.3114	0.3182	0.3050	0.2884	0.1880	0.1510	0.1758	0.2881	**0.1521**
Wine Quality	0.2227	0.2955	0.2958	0.2937	0.2877	0.2575	0.2038	0.1986	0.2529	**0.1408**
Beijing Air Quality	0.2030	0.3481	0.3407	0.3513	0.3279	0.3506	0.2625	0.2042	0.3648	**0.1864**

Based on Fig. 4, our model exhibited strong validity and stability in the four cases, with the MR being 0.2, 0.4, 0.6, and 0.8. Contrarily, other models had apparent fluctuations when the MR changed, particularly at a higher MR. However, the proposed model was not overly affected by changes in MR. On the contrary, MGAIN showed a lower RMSE than other models under different MR. It also exhibited a closer RMSE and a certain stability.

5.4 Prediction performance of MGAIN

We also compared the MGAIN data prediction accuracy after missing data imputation, and it demonstrated the best prediction accuracy. For this purpose, we chose AUC as the performance metric. To be fair, we chose the logistic regression (LR) prediction model for the Default, News, Breast Cancer, and credit dataset with MR of 0.4 and 0.8. The specific results are shown in Tables 8 and 9.

From Table 8, MGAIN is the optimal method for making predictions after missing data imputation, showing the best prediction accuracy. However, the improvement in prediction accuracy was not always significant, even when computational accuracy was greatly improved. For example, on the BreastCancer dataset, the AUC of the data after MGAIN filling was 0.9943. whereas the AUC of the data after the second-best model, PC-GAIN filling, is 0.9913. However, MGAIN improves only 0.003. possibly because there is enough information to predict labels with 0.6 of the observed data. Therefore, we conducted another comparison of the prediction performance with 0.8 missing data, and the results are shown in Table 9.

The results in Table 9 demonstrated that MGAIN was effective at high MR compared with other models. For instance, on the BreastCancer dataset, the AUC of the data after MGAIN filling was 0.0068 higher than that of the data after PC-GAIN filling. The performance gap was even more significant than at MR=0.4. On the News dataset, MGAIN improves its performance over the worst model, Mean, by 0.4244. On the Credit dataset, MGAIN improved its AUC over the second-best model, PC-GAIN, by 0.0311.

In contrast, the latest model, CWGAIN-GP, consistently performed poorly, suggesting its ineffectiveness at handling data with a high MR, which is consistent with the conclusion drawn in Section 5.3 MGAIN outperformed other models in terms of computational performance method. It also exhibited an advantage in prediction performance. Tables 8 and 9 visualize this advantage in Fig. 5, indicating that the proposed model performed better than other models, especially on the Default, News, and Credit datasets. This superiority is attributed to the BreastCancer dataset being smaller by 569, less than other datasets, resulting in less information for the estimation model. Nonetheless, the predictive performance of MGAIN is still better than that of other models.

5.5 Ablation study

To validate the effectiveness of incorporating the least squares loss and dual discriminator in the model, we performed ablation experiments, the specific results of which are shown in Table 10. This experiment was conducted with MR = 0.2.

In Table 10, adding least squares loss and dual discriminator is reasonably practical. The RMSE of GAIN (with least square loss) and GAIN (with dual discriminator) was not as low as the RMSE of GAIN on the News dataset. However, the RMSE of MGAIN was lower than that of GAIN, suggesting that adding these two strategies alone does not fully address gradient vanishing and mode collapse, resulting in a higher RMSE. Since LSGAN solves the gradient vanishing problem but not the mode collapse problem, adding least squares loss or dual discriminator alone does not significantly improve the results. However, D2GAN, alleviates the mode collapse but still faces the gradient vanishing problem. Therefore, this study combines the advantages to allow the proposed MGAIN to alleviate mode collapse and avoid gradient vanishing simultaneously, leading to better overall performance. As a result, MGAIN has high filling accuracy, and good prediction accuracy after filling due to the improved learning ability of the original missing data. Therefore, MGAIN exhibits better performance.

6 Conclusions

The current methods for dealing with missing data are limited. We also observed a lack of effective imputation models, especially when faced with high missing rates. Traditional imputation methods have limitations that can affect model accuracy and stability. Advanced GAN-based imputation methods also face problems such as gradient vanishing and model collapse. Therefore, in this paper, we take the missing data as the research object and construct the MGAIN model. We introduce the least squares loss and dual discriminator to solve these problems. In the empirical analysis, MGAIN was found to reduce the RMSE error by up to 0.2029 when the missing rate was 0.2. In contrast, MGAIN can reduce the RMSE error by up to 0.2166 when the missing rate was 0.8. In addition, we evaluated the prediction of the filled missing data. The experimental results demonstrated that MGAIN outperformed traditional and GAN-based imputation methods. For example, on the Credit dataset, MGAIN improved the prediction performance by 13.65% compared with the

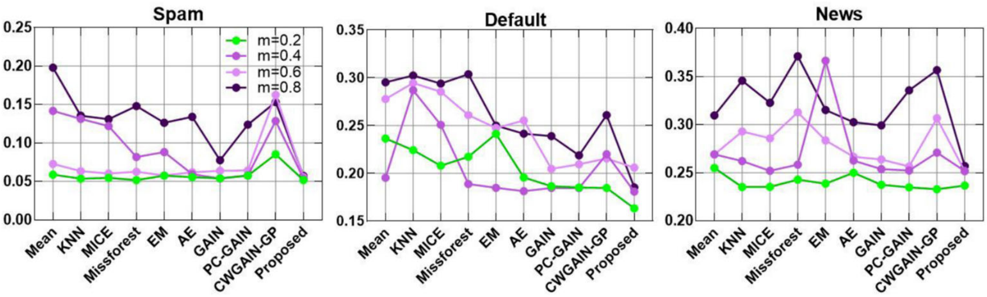


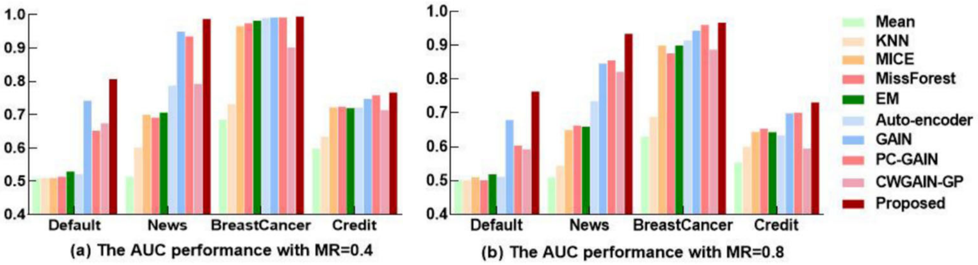
Fig. 4 Analysis of various MR

Table 8 AUC predicted at MR=0.4

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Default	0.5079	0.5096	0.5103	0.5134	0.5289	0.5210	0.7416	0.6518	0.6742	**0.8064**
News	0.5137	0.6015	0.6994	0.6914	0.7055	0.7871	0.9495	0.9348	0.8216	**0.9871**
Breast Cancer	0.6853	0.7311	0.9655	0.9746	0.9820	0.9901	0.9911	0.9913	0.9018	**0.9943**
Credit	0.5984	0.6337	0.7216	0.7235	0.7199	0.7209	0.7468	0.7581	0.7135	**0.7665**

Table 9 AUC predicted at MR=0.8

Datasets	Mean	KNN	MICE	MissForest	EM	Auto-encoder	GAIN	PC-GAIN	CWGAIN-GP	Proposed
Default	0.5000	0.5001	0.5098	0.5007	0.5187	0.5111	0.6786	0.6035	0.5930	**0.7632**
News	0.5094	0.5438	0.6497	0.6635	0.6593	0.7349	0.8461	0.8553	0.7923	**0.9338**
Breast Cancer	0.6301	0.6889	0.8996	0.8759	0.8998	0.9150	0.9437	0.9601	0.8872	**0.9668**
Credit	0.5531	0.5998	0.6438	0.6534	0.6431	0.6339	0.6990	0.7001	0.5947	**0.7312**



(a) The AUC performance with MR=0.4

(b) The AUC performance with MR=0.8

Fig. 5 The AUC performance with MR=0.4(a) and MR=0.8(b)

Table 10 Ablation experiment

Datasets	GAIN	GAIN(with least square loss)	GAIN(with dual discriminator)	Proposed
Spam	0.0540	0.0534	0.0549	**0.0517**
Letter	0.1317	0.1307	0.1314	**0.1214**
Default	0.1864	0.2050	0.2060	**0.1633**
News	0.2375	0.2532	0.2530	**0.2366**
Breast Cancer	0.1008	0.0889	0.0871	**0.0709**
Credit	0.2006	0.2011	0.1984	**0.1956**
MNIST	0.1523	0.1129	0.1141	**0.0969**
Air Quality	0.1146	0.1213	0.0893	**0.0759**
Wine Quality	0.1094	0.1184	0.1084	**0.1036**
Beijing Air Quality	0.0841	0.0734	0.0760	**0.0725**

state-of-the-art CWGAIN-GP model when the missing rate was 0.8. In summary, the proposed model in this paper demonstrates excellent superiority and stability in experimental results. This indicates that MGAIN has important application prospects in real-world scenarios dealing with high missing rates data, which can provide an effective solution for data analysis and prediction tasks.

Author Contributions **Xiwen Qin**: Conceptualization, Writing-original draft, Supervision. **Hongyu Shi**: Methodology, Software, Validation, Writing-review and editing, Conceptualization. **Xiaogang Dong**: Writing editing, Project administration. **Siqi Zhang**: Methodology, Supervision. **Liping Yuan**: Software, Writing-review, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by the National Natural Science Foundation of China(12026430); the Department of Science and Technology of Jilin Province project(20200403182SF). Any opinions, findings, or conclusions in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

Data availability Sources already described in the paper.

Declarations

Conflict of interest/Competing interests The authors declare no conflict of interest.

References

1. Mahmood T, Wittenberg P, Zwetsloot IM, Wang H, Tsui KL (2019) Monitoring data quality for telehealth systems in the presence of missing data. Int J Med Inform 126:156–163
2. Heymans MW, Twisk JW (2022) Handling missing data in clinical research. J Clin Epidemiol 151:185–188
3. Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in r. J Stat Softw 45:1–67
4. Stekhoven DJ, Bühlmann P (2012) Missforest on-parametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118
5. Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11:2287–2322
6. Marchang N, Tripathi R (2020) Knn-st: Exploiting spatio-temporal correlation for missing data inference in environmental crowd sensing. IEEE Sensors J 21(3):3429–3436
7. Jaeger M (2022) The aim and em algorithms for learning from coarse data. J Mach Learn Res 23(62):1–55
8. Ramchandran S, Tikhonov G, Lönnroth O, Tiikkainen P, Lähdesmäki H (2024) Learning conditional variational autoencoders with missing covariates. Pattern Recogn 147:110113
9. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst **27**
10. Lee W, Lee S, Byun J, Kim H, Lee J (2022) Variational cycle-consistent imputation adversarial networks for general missing patterns. Pattern Recogn 129:108720
11. Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23(11):592–596
12. Jiang H, Zhao X, Ma RC, Fan X (2022) Consistent screening procedures in high-dimensional binary classification. Stat Sin 32(1):109–130
13. Valle D, Izbicki R, Leite RV (2023) Quantifying uncertainty in land-use land-cover classification using conformal statistics. Remote Sens Environ 295:113682
14. Volterman W, Davies KF, Balakrishnan N, Ahmadi J (2014) Non-parametric prediction of future order statistics. J Stat Comput Simul 84(3):683–695
15. Thélie E, Aubert D, Gillet N, Hiegel J, Ocvirk P (2023) Topology of reionisation times: Concepts, measurements, and comparisons to gaussian random field predictions. Astron Astrophys 672:184
16. Karageorgiou V, Gill D, Bowden J, Zuber V (2023) Sparse dimensionality reduction approaches in mendelian randomisation with highly correlated exposures. Elife 12:80063
17. Kennedy N, Win TL, Bandyopadhyay A, Kennedy J, Rowe B, McNerney C, Evans J, Hughes K, Bellis MA, Jones A et al (2023) Insights from linking police domestic abuse data and health data in south wales, uk: a linked routine data analysis using decision tree classification. The Lancet Public Health 8(8):629–638
18. Beaulac C, Rosenthal JS (2020) Best: A decision tree algorithm that handles missing values. Comput Stat 35(3):1001–1026
19. Zhao B, Shuai C, Hou P, Qu S, Xu M (2021) Estimation of unit process data for life cycle assessment using a decision tree-based approach. Environ Sci Technol 55(12):8439–8446
20. Behr M, Wang Y, Li X, Yu B (2022) Provable boolean interaction recovery from tree ensemble obtained via random forests. Proc Natl Acad Sci 119(22):2118636119
21. Pinheiro D, Santander-Jimenéz S, Ilic A (2022) Phylomissforest: a random forest framework to construct phylogenetic trees with missing data. BMC genomics 23(1):377
22. Wang S, Qian G, Hopper J (2023) Integrated logistic ridge regression and random forest for phenotype-genotype association analysis in categorical genomic data containing non-ignorable missing values. Appl Math Model 123:1–22
23. Gómez-Méndez I, Joly E (2023) Regression with missing data, a comparison study of techniques based on random forests. J Stat Comput Simul 93(12):1924–1949
24. Zhou Y, Shi J, Stein R, Liu X, Baldassano RN, Forrest CB, Chen Y, Huang J (2023) Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. J Am Med Inform Assoc 30(7):1246–1256
25. Yu L, Li M (2023) A case-based reasoning driven ensemble learning paradigm for financial distress prediction with missing data. Appl Soft Comput 137:110163
26. Schickedanz A, Perales L, Holguin M, Rhone-Collins M, Robinson H, Tehrani N, Smith L, Chung PJ, Szilagyi PG (2023) Clinic-based financial coaching and missed pediatric preventive care: a randomized trial. Pediatr **151**(3)
27. Yu L, Li M, Liu X (2024) A two-stage case-based reasoning driven classification paradigm for financial distress prediction with missing and imbalanced data. Expert Syst Appl 249:123745
28. Sharma A, Hamarneh G (2019) Missing mri pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans Med Imaging 39(4):1170–1183
29. Yoon J, Jordon J, Schaar M (2018) Gain: Missing data imputation using generative adversarial nets. In: International Conference on Machine Learning, pp 5689–5698
30. Awan SE, Bennamoun M, Sohel F, Sanfilippo F, Dwivedi G (2021) Imputation of missing data with class imbalance using conditional generative adversarial networks. Neurocomputing 453:164–171
31. Wang Y, Li D, Li X, Yang M (2021) Pc-gain: Pseudo-label conditional generative adversarial imputation networks for incomplete data. Neural Netw 141:395–403

32. Wang W, Chai Y, Li Y (2022) Gagin: generative adversarial guider imputation network for missing data. Neural Comput & Applic 34(10):7597–7610
33. Wu Z, Ling BWK (2022) Data imputation via conditional generative adversarial network with fuzzy c mean membership based loss term. Appl Intell 52(6):5912–5921
34. Zhao F, Lu Y, Li X, Wang L, Song Y, Fan D, Zhang C, Chen X (2022) Multiple imputation method of missing credit risk assessment data based on generative adversarial networks. Appl Soft Comput 126:109273
35. Al-taezi MA, Wang Y, Zhu P, Hu Q, Al-Badwi A (2024) Improved generative adversarial network with deep metric learning for missing data imputation. Neurocomputing 570:127062
36. Wang Y, Xu X, Hu L, Fan J, Han M (2024) A time series continuous missing values imputation method based on generative adversarial networks. Knowl-Based Syst 283:111215
37. Pham TM, Pandis N, White IR (2022) Missing data, part 2. missing data mechanisms: Missing completely at random, missing at random, missing not at random, and why they matter. Am J Orthod Dentofac Orthop 162(1):138–139
38. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2794–2802
39. Nguyen T, Le T, Vu H, Phung D (2017) Dual discriminator generative adversarial nets. Adv Neural Inf Process Syst **30**
40. Stefenon SF, Seman LO, Mariani VC, Coelho LdS (2023) Aggregating prophet and seasonal trend decomposition for time series forecasting of italian electricity spot prices. Energ 16(3):1371

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xiwen Qin is now a professor and doctoral supervisor at the School of Mathematics and Statistics at Changchun University of Technology. His current research directions include machine learning, big data analysis and intelligent optimization, and time series analysis. He has published more than 35 papers in academic journals and 1 monograph.

Hongyu Shi is currently pursuing her PhD degree at Changchun University of Technology. Her research interests are in the theory of generative adversarial networks and their downstream tasks. Examples include data augmentation, missing data imputation, and time series prediction.

Xiaogang Dong is now a professor and doctoral supervisor at the School of Mathematics and Statistics at Changchun University of Technology. He has been engaged in research in financial big data analysis, time series analysis, and machine learning. In the past three years, he has published more than 45 academic papers.

Siqi Zhang is currently a PhD student in Statistics at Changchun University of Technology. She has published 2 SCI papers. Her main research interests are imbalanced data, feature selection, swarm intelligent optimization algorithms and federated learning.

Liping Yuan is currently working towards her Doctor's degree in statistics at Changchun University of Technology. Her main research interests include swarm intelligence optimization algorithms, time series prediction, and mixed frequency data analysis.