Multiple Imputation by Generative Adversarial Networks for Classification with Incomplete Data

Bao Ngoc Vi
Data Science Research Group
Faculty of Information Technology
Le Qui Don Technical University
Hanoi, Vietnam
ngocvb@lqdtu.edu.vn

Dinh Tan Nguyen
AI Academy Vietnam
Hanoi, Vietnam
tommy@aiacademy.edu.vn

Cao Truong Tran
Data Science Research Group
Faculty of Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
truongct@lqdtu.edu.vn

Huu Phuc Ngo
Data Science Research Group
Faculty of Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
phucnh@lqdtu.edu.vn

Chi Cong Nguyen
Data Science Research Group
Faculty of Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
congnc@lqdtu.edu.vn

Hai-Hong Phan
Data Science Research Group
Faculty of Information Technology
Le Quy Don Technical University
Hanoi, Vietnam
hongpth@lqdtu.edu.vn

***Abstract*—Missing values present as the most common problem in real-world data science. Inadequate treatment of missing values could often result in mass errors. Hence missing values should be managed conscientiously for classification. Generative Adversarial Networks (GANs) have been applied for imputing missing values in most recent years. This paper proposes a multiple imputation method to estimate missing values for classification through the integration of GAN and ensemble learning. Our propose method MIGAN utilises GAN to generate different training observations which are then used to conduct ensemble classifiers for classification with missing data. We conducted our experiments examine MIGAN on various data sets as well as comparing MIGAN with the state-of-the-art imputation methods. The experimental results show significant results, which highlights the accuracy of MIGAN in classifying the missing data.**

***Index Terms*—missing data, incomplete data, imputation, generative adversarial network, ensemble learning**

I. Introduction

Missing values where the values of some features are unknown and have presented as one of the common issues in many real-world datasets. For instance, about 45% of the UCI machine learning repository [1] often encounters with the missing values [2]. There are different causes of missing values. For example, in social surveys, when respondents tended to deny to reply in specific questions, the collected datasets will be incomplete [13]. Furthermore, medical datasets usually contain a huge number of missing values since it is extreme rare to achieve 100% of task completion on every patient [4]. Missing values are causing significant problems in classification. One of the most serious problems occurs when majority of classification algorithms failed to work on incomplete datasets [2]. Another issue is that missing values often lead to mass classification error due to insufficient information during either the training or application processes [5].

One of the most common approaches for classification with missing data is via the usage of imputation methods to transform incomplete data into complete data. For example, mean imputation fills each missing field with the average of the complete values [2]. Although imputation is the most common way of classification with missing data, unable to carefully imputing missing values could also lead to mass classification errors. Therefore, it is essential to propose robust imputation methods of classification with missing data.

A generative adversarial network (GAN) is a machine learning model in which two neural networks compete with each other in a zero-sum game. GANs have been successfully applied in many applications such as human image synthesis, improving astronomical images and inpainting photographs [3]. Recently, GANs has also been applied to imputing missing values such as GAIN [15], MisGan [9], and GAMIN [16].

This paper proposes a key method to accurately impute missing values for classification. The proposed method uses GANs to generate multiple complete data sets from an incomplete data set. After that, a set of classifiers is build on these complete datasets. We aim to show:

1) The proposed method can be more accurate than the other existing GANs-based imputation methods;
2) The proposed method can be more accurate than the other existing traditional imputation methods.

II. Related Work

A. Missing Data

Missing data has been a pervasive topic in the empirical research. Data science studies have constantly been encounter-

978-1-6654-0435-8/21/$31.00 ©2021 IEEE

ing with many types of missing data; they are MCAR, MAR and MNAR respectively and could be categorized based on the mechanisms of missing data article. Missing Completely At Random (MCAR) occur when the missingness is not associated with its hypothetical value, the values of other variables or sets of observed records. Missing At Random (MAR) happens when the propensity of any missing data points are not related to the specific missing values; where that specific missing values are expected to be obtained. However it also depends on few of the observed data. Furthermore, the ultimate type of Missing Data is called Missing Not At Random (MNAR), in which the missing data points depend on both of the hypothetical values and other specific variable's values.

To avoid bringing negative effects in the validity of the trials of our experiments, many methods and approaches have been provided to manipulate the missing data. One of the common way is the deletion method and is often utilised by data scientists, deleting incomplete features or incomplete samples during the imputation process. By omitting partial or entire missing records, the remaining data will continue to be utilised as the original data. However, when the missing rate is not at its minimal (normally 5% of the whole data), the deletion method will not be the best approach as it can lead to the generation of deficient outputs [11]. Furthermore, imputation methods which substitute missing values with plausible values are overall a better approach towards missing data.

B. Imputation methods

Imputation method is the most common strategy in handling missing data [8]. The concept is filled up with a missing value with a "well-calculated estimate". By taking advantages from statistical analysis and mathematical models, varieties of approaches can be designated as maximum likelihood, expectation-maximization, regression imputation, multiple imputations, sensitivity analysis [8], [10].

Machine learning has been applied to impute missing values. Decision trees and k-nearest neighbors are often used to impute missing values [6], [11]. With the improvement of machine learning algorithms, machine learning-based imputation methods tended to sustain its accuracy in many experiments. However, the drawbacks of these methods are highly calculated on the quality of data. If a dataset is relatively small or high-dimensional, some of the models are seemingly less sufficient. Thus, the selective model should be considering with distinctive datasets [11].

Imputation methods can be categorised into single imputation methods [11] and multiple imputation methods [12]. While a single imputation method just creates one complete data set from an incomplete data set, a multiple imputation method is able to generate multiple datasets from an incomplete dataset. Single imputation tends to rely on specific assumptions of missing values rather than the type of missing data [12]. These assumptions are not applicable or identical, and they often lead to bias results. Multiple imputation methods are valid general method for reducing the bias of imputation. In general, multiple imputation methods improve the validity of process, its procedure adds random value to restore randomness loss. By reducing randomness, the statistical analysis on distribution will be more appropriate. Multiple imputation is also more flexible and is be applied in a wide variety of scenarios [11].

C. Generative Adversarial Networks for Data Imputation

Generative adversarial networks (GANs) are the subject of debate in most recent years. GANs are comprised of a generator and a discriminator, which are trained in an adversarial way. These two models are usually implemented by the neural networks. GANs have been successfully applied to various fields such as image processing and computer vision, natural language processing, and medicine [3] . Inspire of the success of GANs in data synthesis, researchers have applied GANs to data imputation. The GAN-based data imputation methods firstly proposed to image completion [17], [18]. However, these models applied in image inpainting only. Recently, there are some publishes on data imputation in general, such as GAIN [15], MisGan [9], and GAMIN [16].

In Generative Adversarial Imputation Nets (GAIN) model [15], the generator is considered as a imputer and the discriminator tries to discriminate whether each components of an input had already been imputed or not. The algorithm is accurate in low dimensional datasets with low missing rate. It also works at MNIST of 50% missing rate. However it converges to zero imputation or mean imputation at higher missing rates.

On the other hand, MisGAN model [9] works better for datasets at highly missing rate. The proposed method consists of a GAN architecture for missing dataset and an imputer using it. This GAN architecture is comprised of two generators pairs for mask and data respectively. The data generator is responsible for generating fake complete data and the mask generator is responsible for generating the fake mask which indicates which component is missed. Then, the fake complete data is masked with the fake mask to generate the fake missing data. The data discriminator tries to distinguish real missing data from fake missing data. Another pair of generator and discriminator are used for data imputation. The imputation generator imputes missing data to fool the corresponding discriminator which distinguishes imputed data from fake complete data.

The generative adversarial multiple imputation network (GAMIN) [16] is proposed for multiple imputation for highly missing data. This model is motivated by the MisGAN but there are several changes. Firstly, the imputation architecture is changed to make the data generator be directly included in imputation process. Secondly, a novel confidence prediction and top-k imputation is introduced. Finally, GAMIN is trained using new loss functions considering the confidence. Moreover, the mask generation is replaced by the mask of missing data. The unconditional data generation is replaced with the conditional generation using the missing data .

III. The Proposed Method

Most of GAN-based imputations are stochastic due to the presence of random noise which is fed to generator. Moreover, multiple imputations can be obtained by repeating the single imputation process. Therefore, in this paper, we propose a multiple imputation method to estimate missing values for classification by integrating GAN and ensemble learning (namely MIGAN). There are two main steps in our model. Firstly, the GAN-based imputation is trained. Secondly, the trained generator is used to generate different complete data sets which are used to build an ensemble classifiers.

A. GAN-based Imputation

Our GAN-based imputation is developed based on GAIN [15]. While GAIN is a single imputation method, the proposed method is a multiple imputation method by repeating GAIN multiple times to conduct multiple complete data sets from an incomplete data set.

1) Generator: Suppose that $\mathbf{X} = (X_1, X_2, ..., X_d)$ is a missing data in d-dimensional space, $\mathbf{M} = (M_1, M_2, ..., M_d)$ is a mask vector indicating which components of $\mathbf{X}$ are observed, that is:

$$M_i = \begin{cases} 1 \text{ if } X_i \text{ is observed} \\ 0 \text{ if } X_i \text{ is unobserved} \end{cases}$$

The generator G take $\mathbf{X}$, $\mathbf{M}$ and a d-dimension noise variable $\mathbf{Z}$ as input and outputs a vector of imputations $\bar{\mathbf{X}}$:

$$\bar{\mathbf{X}} = G(\mathbf{X}, \mathbf{M}, (\mathbf{1} - \mathbf{M}) \odot \mathbf{Z})$$

where $\mathbf{Z}$ is sampled from a known distribution such as normal distribution, and $\odot$ is element-wise multiplication.

G outputs a value for every component as long as its value is observed. Therefore the final completed data vector is obtained by replacing missing value in $\mathbf{X}$ with the corresponding value of $\bar{\mathbf{X}}$, that is:

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

With one pair of $\mathbf{X}$ and $\mathbf{M}$, we can repeat sampling $\mathbf{Z}$ multiple times to obtain multiple imputations of missing data $\hat{\mathbf{X}}$.

2) Discriminator: The discriminator attempts to distinguish which components of completed vector $\hat{\mathbf{X}}$ are real (observed) or fake (imputed). That is, it try to predicting the mask vector $\mathbf{M}$. In [15], authors showed that if the discriminator D is not provided "enough" information about M, G could reproduce several populations that would all be optimal with respect to D. Thus it is necessary to introduce a hint mechanism. A hint mechanism is a random variable $\mathbf{H}$ and contains some information about M to guarantee that the generator learns the desired distribution. $\mathbf{H}$ is defined as below:

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

where $\mathbf{B} = (B_1, B_2, ..., B_d) \in \{0, 1\}^d$ is defined by first sampling k from $\{1, ..., d\}$ uniformly at random and then setting:

$$B_j = \begin{cases} 1 \text{ if } j \neq k \\ 0 \text{ if } j = k \end{cases}$$

$\hat{\mathbf{X}}$ and $\mathbf{H}$ are fed to D to predict which components are real or imputed.

3) Training objective: D is trained to maximize the probability of correctly predicting $\mathbf{M}$ while G is trained to minimize the probability of D predicting $\mathbf{M}$. Moreover, in training G, we not only ensure that estimate successfully missing components to fool D but also ensure that the values outputted by G for observed components are close to those actually observed. Then the training objective is defined as below:

$$\min_G \max_D \mathbb{E}\left[L(\mathbf{M}, D(\hat{\mathbf{X}}, \mathbf{H}), \mathbf{B}) + L_{\mathbf{M}}(\mathbf{X}, \bar{\mathbf{X}})\right] \tag{2}$$

where

$$L(\mathbf{M}, \hat{\mathbf{M}}, \mathbf{B}) = \sum_{i:B_i=0} \left[M_i log(\hat{M}_i) + (1 - M_i) log(1 - \hat{M}_i)\right]$$

and

$$L_{\mathbf{M}}(\mathbf{X}, \bar{\mathbf{X}}) = \sum_{i=1}^{d} (X_i - \bar{X}_i)^2$$

B. GAN-based Multiple Imputation and Ensemble Learning

This step consists of two phases: the training phase and the application phase. In the training phase, the generator, which is used as a multiple imputer, integrates with ensemble learning to build a set of classifiers. Subsequently, in the application process, the multiple imputer and the set of classifiers are used together to classify a new incomplete sample.

In the training phase, a training incomplete dataset is fed to G together with multiple set of random noises $\mathbf{Z}$ to build a set of imputed datasets. Then, each imputed dataset is independently used to train a classifier. Consequently, a set of classifiers are obtained.

In the application phase, if a sample which needs to be classified has missing values, it is firstly fed to G together with a set of random vectors z to generate a set of imputed samples. Then, the average of these imputed samples is calculated to get an unique imputed sample. After that, this average sample is fed to each classifier to obtain a set of predicted classes. The predicted class is a class with the highest majority of votes, that is the class which had the highest probability of being predicted by each of the classifiers. On the other hand, if a sample which needs to be classified is complete, the imputation step can be omitted.

Our model exploits the stochastic characteristic of GAN-based imputation to build multiple imputation and uses it to build a set of diverse imputed datasets. As a result, diverse classifiers are able to be constructed, which makes the ensemble classifier efficient.

IV. Experiment Design

A. Experiment methodology

A series of experiments are conducted to validate the performance of the Multiple Imputation by Generative Adversarial Networks for Classification (MIGAN). Quantitatively, UCI datasets [14], which include the Spam, Breast, Banknote and Parkinsons, are used to evaluate the performance of all the

tested methods following by the comparisons between basic and advanced methods of missing data imputation. These are all complete datasets, though to make missing datasets we randomly remove some of the data points (MCAR). The missing rate is the percentage of data points removed.

Additionally, a comparison between MIGAN at different missing rates is included to test the stability of MIGAN. In our last experiments, the comparisons between MIGAN against other imputation algorithms are also conducted. Each experiment is run for 30 times to maintain the robustness of all tested methods. And the performance metrics are reported with accuracy depending based on the structure of datasets [7]. Mean and standard deviations are shown to identify the qualitative results across all of the imputation methods.

B. Experiment Settings

Several imputaion methods are used to compare with MIGAN. Traditional imputation methods included mean imputation and *kNN* imputation. MICE represented for Multiple imputation. In MICE, the estimator variate is decision tree regressor. With regards to classification task, the imputed data is divided to 70% for training and 30% for testing. Decision tree (DT) and multi-layer perceptron (MLP) are used to classify imputed datasets.

V. Results and Discussion

A. Comparison of MIGAN and GAIN

In this testing section, we would like to validate the performance of MIGAN using all mentioned UCI datasets. This comparison is made with the purpose of comparing MIGAN with GAIN. The settings of all imputation methods are designed to follow the experiment setting section. The missing rate is changing from 10% to 50% for every datasets and every classifiers. Furthermore, we conducted each test for 30 times to calculate the mean and standard variation of the output.

Table I shows the results of all run-time. In which the MIGAN outperforms GAIN in majority of the run-time from 1% to 5%. This is probably due to its more advanced analytic process and imputation process, which brings better results in predicting the label in classification tasks.

B. Missing Rate Experiment

To better evaluate MIGAN, several testings in varying the missing rate of data are conducted, ranging from 10% to 50% for all datasets (UCI datasets). Figure 1 and 2 shows the performance (accuracy metrics) of MIGAN with the other methods. Figure 2 shows the outputs of the imputed data from experiments aligned with MLP classifier. The accuracy metric is recruited in this method.

Figure 1 and 2 affirm that the MIGAN tends to be more robust than other methods when improving the missing rate. Figure 1a, 1c, 1d, 2b, 2c, 2d highlight that the quality of imputed data against the high level of missing rate. MIGAN outperforms than the others in Decision Tree tests, which are shown in Figure 1a, 1c, 1d. As information from observed data decreases, the imputation methods are able to demonstrate the ability of learning the remaining information when imputing the missing data points. Similar results for MLP tests can also be seen in Figure 2a, 2b, 2c, 2d. Thus MIGAN had significantly pinpointed its imputation quality in dealing with a remarkable missing datasets comparing to the rest of the other tested methods.

C. Comparison of MIGAN and Others

In this experiment, we used all imputation methods including MIGAN, GAIN, Mean, KNN, and MICE in the testing of both Decision Tree and MLP classifications. The missing rate is set from 10% to 50% for these datasets aligning with other similar settings. The result with missing rate 20% had been illustrated in Table II. The visualisation of other results are revealed in Figure 1 from the missing rate varying from 10%, 20%, 30% , 40% and 50%. This experiment aimed to portray the differences between all imputation methods with the same input data.

In this run, MIGAN have shown significant performance with respect to the accuracy of post-imputation prediction. The standard variations of MIGAN from Table II are however relatively smaller than other methods in all the testings. The gap between MIGAN and others is wider in Decision Tree task, but smaller in MLP due to the adaptive learning of MLP, which yield the required decision function directly via training. MICE in Figure 1d had shown with better result in some specific dataset like Parkinsons dataset, however MIGAN still performed more advanced comprehensively towards the rest of the testings.

VI. Conclusion

In this paper, we proposed a new approach called the "Multiple Imputation by Generative Adversarial Networks for Classification (MIGAN)" to perform multiple imputation to estimate the missing values in incomplete data sets. This novel framework is advanced by integrating GAN frameworks, multiple imputation concepts and ensemble learning paradigm in significant errors and bias reduction during the process of imputation of missing data. Our experiments highlighted multiple positive performance of MIGAN against other imputation methods with higher accuracy.

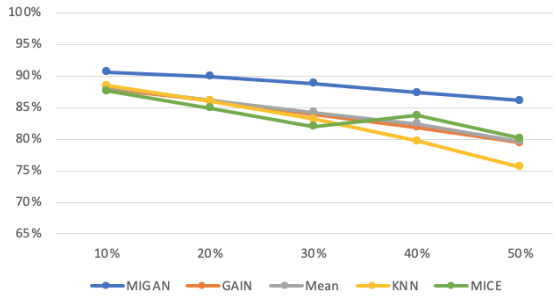
In the coming future, we could focus on reinforcing the stability of MIGAN and to improve the quality and efficacy of imputing incomplete data.

References

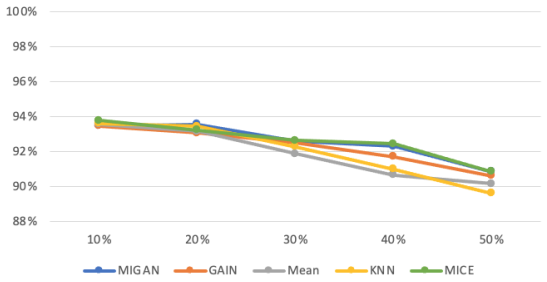
[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

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

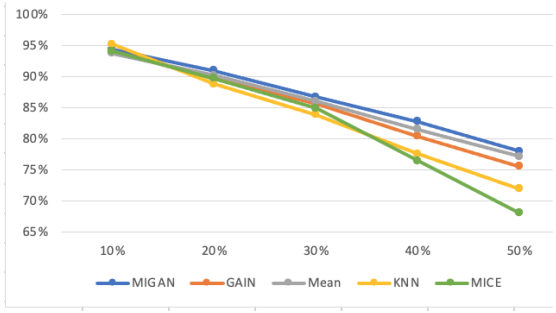
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.



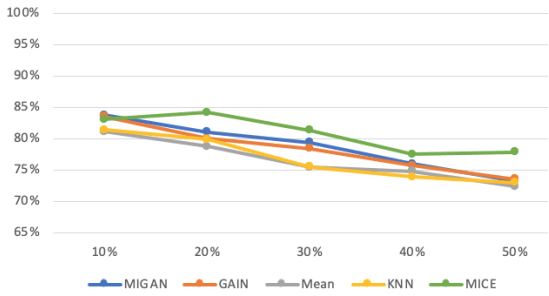
(a) Spam Dataset with Decision Tree Classifier



(b) Breast Dataset with Decision Tree Classifier

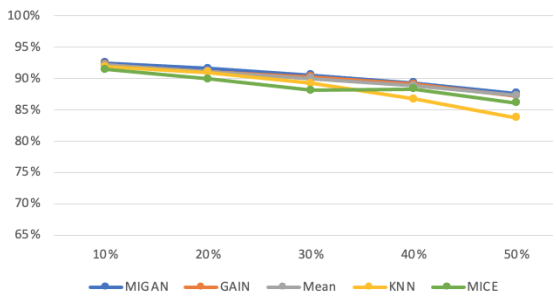


(c) Banknote Dataset with Decision Tree Classifier

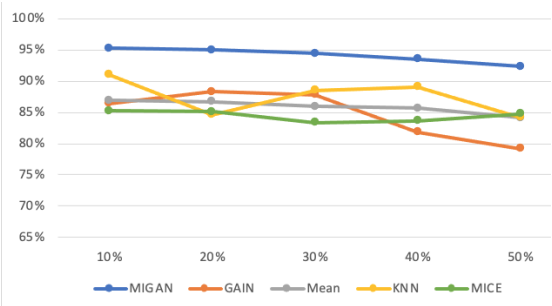


(d) Parkinsons Dataset with Decision Tree Classifier

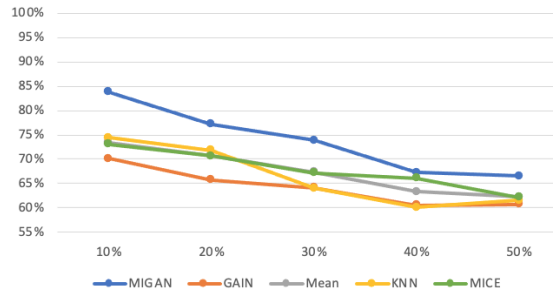
Fig. 1: UCI Datasets over missing rates testing using Decision Tree



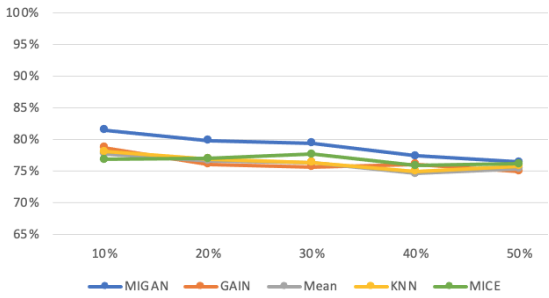
(a) Spam Dataset with MLP Classifier



(b) Breast Dataset with MLP Classifier



(c) Banknote Dataset with MLP Classifier



(d) Parkinsons Dataset with MLP Classifier

Fig. 2: UCI Datasets over missing rates using MLP

TABLE I: MIGAN vs GAIN imputation method

Missing Rate	Classifier	Imputation %	Spam	Breast	Banknote	Parkinsons
10%	DT	MIGAN	90.61 ± 0.90	93.46 ± 1.11	94.37 ± 1.38	83.73 ± 4.38
	DT	GAIN	87.99 ± 0.62	93.46 ± 1.23	94.09 ± 1.25	83.56 ± 4.59
	MLP	MIGAN	92.45 ± 0.40	95.24 ± 0.85	83.79 ± 9.59	81.53 ± 1.91
	MLP	GAIN	92.15 ± 0.70	86.41 ± 18.49	70.10 ± 16.71	78.70 ± 3.55
20%	DT	MIGAN	89.93 ± 0.80	93.56 ± 1.22	90.96 ± 1.42	81.02 ± 4.66
	DT	GAIN	86.08 ± 0.92	93.06 ± 1.61	89.74 ± 1.76	80.00 ± 4.98
	MLP	MIGAN	91.55 ± 0.49	94.98 ± 0.73	77.14 ± 7.73	79.83 ± 2.93
	MLP	GAIN	91.09 ± 0.88	88.33 ± 15.87	65.78 ± 14.03	76.05 ± 4.07
30%	DT	MIGAN	88.80 ± 0.75	92.59 ± 1.39	86.73 ± 1.70	79.38 ± 4.67
	DT	GAIN	83.86 ± 1.04	92.49 ± 1.50	85.63 ± 1.83	78.36 ± 4.72
	MLP	MIGAN	90.54 ± 0.70	94.40 ± 0.90	73.86 ± 7.17	79.44 ± 3.04
	MLP	GAIN	90.27 ± 0.84	87.78 ± 13.38	64.12 ± 12.74	75.65 ± 3.46
40%	DT	MIGAN	87.35 ± 0.76	92.30 ± 1.29	82.73 ± 1.92	75.99 ± 5.45
	DT	GAIN	81.85 ± 1.12	91.71 ± 1.31	80.40 ± 2.11	75.65 ± 5.96
	MLP	MIGAN	89.25 ± 0.75	93.54 ± 1.48	67.29 ± 6.32	77.40 ± 2.86
	MLP	GAIN	89.06 ± 0.78	81.86 ± 18.61	60.52 ± 10.86	76.05 ± 3.01
50%	DT	MIGAN	86.09 ± 0.86	90.84 ± 2.24	77.97 ± 2.58	73.17 ± 5.02
	DT	GAIN	79.38 ± 1.12	90.60 ± 2.22	75.54 ± 2.12	73.50 ± 5.29
	MLP	MIGAN	87.58 ± 0.75	92.32 ± 1.65	66.53 ± 4.27	76.39 ± 2.18
	MLP	GAIN	87.21 ± 1.15	79.19 ± 21.97	60.74 ± 8.71	74.97 ± 2.80

TABLE II: MIGAN vs other imputation methods

Classifiers	Imputation %	Spam	Breast	Banknote	Parkinsons
Decision Tree	MIGAN	**89.93 ± 0.80**	**93.56 ± 1.22**	**90.96 ± 1.42**	**81.02 ± 4.66**
	GAIN	86.08 ± 0.92	93.06 ± 1.61	89.74 ± 1.76	80.00 ± 4.98
	Mean	86.12 ± 1.09	93.16 ± 1.38	90.20 ± 1.20	78.76 ± 4.38
	KNN	86.03 ± 0.80	93.44 ± 1.26	88.78 ± 1.74	79.83 ± 5.27
	MICE	84.91 ± 1.05	93.22 ± 1.60	89.72 ± 1.52	84.18 ± 5.06
MLP Regression	MIGAN	**91.55 ± 0.49**	**94.98 ± 0.73**	**77.14 ± 7.73**	**79.83 ± 2.93**
	GAIN	91.09 ± 0.88	88.33 ± 15.87	65.78 ± 14.03	76.05 ± 4.07
	Mean	91.11 ± 0.77	86.67 ± 16.64	70.76 ± 15.00	76.61 ± 3.57
	KNN	90.90 ± 0.77	84.64 ± 19.04	71.85 ± 15.93	76.89 ± 3.43
	MICE	89.89 ± 0.85	85.13 ± 19.33	70.63 ± 15.46	77.01 ± 4.79

[4] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50:105–115, 2010.

[5] J. Luengo, S. García, and F. Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems*, pages 1–32, 2012.

[6] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[7] M. R. Segal. Machine learning benchmarks and random forest regression. *IEEE Symposiumon Security and Privacy*, 2004.

[8] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338(jun29 1), 2009.

[9] B. M. M. Steven Cheng-Xian Li, Bo Jiang. Misgan: Learning from incomplete data with generative adversarial networks. *arXiv preprint arXiv:1902.09599*, 2019.

[10] D. Talwar, A. Mongia, D. Sengupta, and A. Majumdar. Autoimpute: Autoencoder based imputation of single-cell rna-seq data. *Scientific Reports*, 8(1), 2018.

[11] C. T. Tran, M. Zhang, and P. Andreae. Multiple imputation for missing data using genetic programming. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, GECCO '15, page 583–590, New York, NY, USA, 2015. Association for Computing Machinery.

[12] C. T. Tran, M. Zhang, P. Andreae, and B. Xue. Multiple imputation and genetic programming for classification with incomplete data. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 521–528, 2017.

[13] N. Tsikriktsis. A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, 24:53–62, 2005.

[14] I. Yeh. Uci machine learning repository: Data set. *Irvine: University of California*, 2007.

[15] J. Yoon, J. Jordon, and M. van der Schaar. GAIN: missing data imputation using generative adversarial nets. *CoRR*, abs/1806.02920, 2018.

[16] S. Yoon and S. Sull. Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8464, 2020.

[17] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.