Mixed Data Imputation Using Generative Adversarial Networks

WASIF KHAN, NAZAR ZAKI, AMIR AHMAD, MOHAMMAD MEHEDY MASUD, LUQMAN ALI, NASLOON ALI, AND LUAI A. AHMED

1Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates
2Department of Information Systems and Security, College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates
3College of Medicine and Health Sciences, Institute of Public Health, United Arab Emirates University, Al Ain, United Arab Emirates
4Zayed Centre for Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding author: Nazar Zaki (nzaki@uaeu.ac.ae)

This work was supported by a grant from Zayed Center for Health Sciences, United Arab Emirates University (31R239).

ABSTRACT

Missing values are common in real-world datasets and pose a significant challenge to the performance of statistical and machine learning models. Generally, missing values are imputed using statistical methods, such as the mean, median, mode, or machine learning approaches. These approaches are limited to either numerical or categorical data. Imputation in mixed datasets that contain both numerical and categorical attributes is challenging and has received little attention. Machine learning-based imputation algorithms usually require a large amount of training data. However, obtaining such data is difficult. Furthermore, no considerable work has been conducted in the literature that focuses on the effects of the training and testing size with increasing amounts of missing data. To address this gap, we proposed that increasing the amount of training data will improve imputation performance. We first used generative adversarial network (GAN) methods to increase the amount of training data. We considered two state-of-the-art GANs (tabular and conditional tabular) to add synthetic samples using observed data with different synthetic sample ratios. We then used three state-of-the-art imputation models that can handle mixed data: MissForest, multivariate imputation by chained equations, and denoising auto encoder (DAE). We proposed robust experimental setups on four publicly available datasets with different training-testing data divisions that have increasing missingness ratios. Extensive experimental results show that incorporating synthetic samples with training data achieves better performance compared to the baseline methods for mixed data imputation in both categorical and numerical variables, especially for large missingness ratios.

INDEX TERMS

Mixed data imputation, missing data, GANs, miss forest, MICE, denoising auto encoders.

I. INTRODUCTION

Missing information is one of the most common issues in real-world datasets and pose a significant challenge to the performance of statistical and machine learning models. Generally, missing values are imputed using statistical methods, such as the mean, median, mode, or machine learning approaches. These approaches are limited to either numerical or categorical data. Imputation in mixed datasets that contain both numerical and categorical attributes is challenging and has received little attention. Machine learning-based imputation algorithms usually require a large amount of training data. However, obtaining such data is difficult. Furthermore, no considerable work has been conducted in the literature that focuses on the effects of the training and testing size with increasing amounts of missing data. To address this gap, we proposed that increasing the amount of training data will improve imputation performance. We first used generative adversarial network (GAN) methods to increase the amount of training data. We considered two state-of-the-art GANs (tabular and conditional tabular) to add synthetic samples using observed data with different synthetic sample ratios. We then used three state-of-the-art imputation models that can handle mixed data: MissForest, multivariate imputation by chained equations, and denoising auto encoder (DAE). We proposed robust experimental setups on four publicly available datasets with different training-testing data divisions that have increasing missingness ratios. Extensive experimental results show that incorporating synthetic samples with training data achieves better performance compared to the baseline methods for mixed data imputation in both categorical and numerical variables, especially for large missingness ratios.
finds maximum likelihood estimates, are used for better missing data imputation. However, one of the major limitations of the EM algorithm is that it does not provide standard errors as an automatic part of the process [8].

Machine Learning (ML)-based algorithms are used to predict missing values based on the information available in the dataset [1], [10], [11]. The ML algorithms, such as multilayer perceptron, K-nearest neighbors, and support vector machines, use the information from the available complete data to estimate the missing values with high precision. More recently, deep learning-based algorithms have been proposed to automatically learn the complex inter-variable association and estimate more realistic values [12]. One of the deep learning-based algorithms is known as denoising autoencoders (DAEs) [13] which are designed to reconstruct an original input from noisy data. The DAE consists of an encoder and decoder where the dataset with missing entries serves as an input to the encoder and missing entries serve as the noise. The decoder reconstructs the complete dataset by predicting the missing values. Details of the DAE can be found in section x. However, the DAE requires full data and initial guesses are required before imputation [14], [15], such as the work of Ranjit et al. [15] using 0 as the initialization.

Tabular data can be grouped into three categories: numeric data, categorical (nominal) data, and mixed data. An example of a mixed dataset is shown in Table 1, in which the age, weight, height, and Body Mass Index (BMI) are numeric data, while smoking, education, and the diagnosis of diabetes are categorical (nominal) data. Table 1 shows that some of the entries are missing (remain empty with no response). Mixed datasets are difficult due to problems in finding the similarity between two points (categorical and numerical) [16], [17]. Similarly, imputation in numerical values is easier; thus, multiple algorithms have been proposed with great success [4], [14], [19], [20], [21], [22], [23], [24], [25]. However, these algorithms cannot work on most real datasets, such as genomic data, industrial data, and equipment maintenance, because they are often mixed datasets [26]. Imputing mixed datasets is a challenge in many ways, including measuring the relationships between the categorical and numerical instances, hybrid estimators for numerical and categorical data, and considering the attributes of purely categorical or numerical datasets [26], [27]. Furthermore, it is difficult to calculate the generalized similarity measures between features in mixed datasets [28], [29], [16]. The missingness in data occurs due to various reasons, which can be categorized as three types of missingness mechanisms [5], [18]. The first is missingness at random (MAR) in which the missingness depends on the observed data. In the missing not-at-random (MNAR) mechanism, the missingness depends on the missing attributes. In the missing completely at random (MCAR) mechanism, the missingness does not depend on any missing or observed attributes, i.e., the missingness cannot be explained by the data.

Deep learning-based algorithms usually require sufficient training data, which is one of the primary challenges for such models. Although a large amount of data is generated in everyday life, such as in hospitals, retail, education, transportation, and energy, privacy and copyright issues only allow a small proportion of the datasets to be publicly available [30]. In addition, publicly available datasets usually contain a limited number of samples with missing values. Furthermore, obtaining a large amount of data is difficult because it relies on humans as the gold standard, and may require verifications with multiple sources, which is expensive, time-consuming, and subjective [31]. Alternatively, data oversampling is used where artificial samples are created from the original training data to enlarge the training dataset. Adding synthetic samples to the data increases the generalization and avoids overfitting [32]. Data augmentation is common for images [31], [32] [33], [34], [35], [36]; however, it has received less attention for tabular data augmentation [37] [38]. In images, artificial samples can be created by rotating, flipping, zooming, shearing, and adding noise to images. However, generative adversarial networks (GANs), as introduced by Goodfellow et al. [39], are a more sophisticated method to create realistic synthetic samples. GANs are used for artificial data generation, which consists of two neural networks: the generator and the discriminator. The generator is used to generate synthesized artificial samples with required variations from the input data distribution while the discriminator differentiates between the samples created by the generator or from the input data. Multiple modified versions of the GAN are introduced to synthesize new data for both images and tabular data. This paper uses two versions of GANs (TGAN [40] and CTGAN [41]) to oversample tabular data.

The major contributions of this paper are summarized as follows:

- We propose a novel approach that incorporates synthetic samples with training data to improve the imputation performance using two different GANs to synthesize tabular training data.
- We provide extensive experiments on four mixed datasets with each set divided into different training-testing sets of 80-20, 60-40, 40-60, and 20-80 with an increasing missing ratio of 10%, 30%, 50%, and 70%.
- We provide a detailed performance evaluation for three missing data imputation methods using mixed datasets.

The rest of the paper is organized as follows. Section II provides an overview of related works. The methodology of the proposed work is described in Section III. Section IV gives the experiments and presents the results. Section V provides the discussion. Finally, the conclusion is presented in Section VI.

II. RELATED WORKS

This section provides a brief overview of the studies performed for missing data imputation. We first describe studies that used ML models and deep learning for missing data imputation with numerical or mixed datasets. Imputation of missing values is common to reduce bias due to missingness and aids in improving the statistical and ML-based
analyses [42]. The simplest method to impute missing values is using the mean for numerical values and mode for categorical data [43]; however, these methods ignore the relationships and underestimate variance with the other variables [44]. The ML-based techniques create a prediction model from the observed data to determine the missing values [25]. Missing value imputation for numerical datasets is common [14], [19], [20], [21], [22], [23], [4], [24], [25] and include the nearest neighbor [22], LLS [45], SVD [22], and ensemble [26] approaches. Deep learning-based DAE [14] is also used for numerical data imputation.

Imputing missing values in a mixed dataset is a challenge and has received little attention. Some of the work that uses missing data imputation for mixed datasets are briefly described as follows. Methods such as the MissForest [46] use an RF classifier for missing value imputation, while MICE [47] algorithms are used for mixed data imputations (see further details in the methodology section). Alsaber et al. [46] used MissForest imputation on air quality data obtained from the Kuwait Environmental Public Authority from 2012–2018. The authors showed that introducing multiple missingness ratios of 5%, 10%, 20%, and 40% in the dataset gave MissForest a better prediction accuracy compared to other methods such as KNN, MICE, Amelia, and missCompare. Aleryani et al. [47] used ensembles of MICE and EMB using bagging and stacking ensembles for data imputation. The evaluation was performed on 20 datasets, which shows the ensemble methods outperform individual imputation techniques.

Awawdeh et al. 2022 [48] proposed a genetic algorithm (GA) based simultaneously on missing data imputation and feature selection. The proposed GA-based method divided the dataset into observed and missing data. The observed data were used to train the model, and a subset of features that contain missing values was imputed with the initial value (mean or mode), which was used as a test set. Then, the imputed instances were used as a test set to evaluate the performance of the trained model, which was trained based on the complete data. The process of imputing data in the test set was repeated based on the GA-based stopping criteria until the optimal subset of the imputed dataset was obtained. The authors evaluated their method on 12 different datasets and compared their method with other imputation techniques.

Several studies have used deep learning-based algorithms for missing data imputation on mixed data as well as numerical data. Lall et al. [15] proposed deep learning based on multiple imputations using DAEs for mixed datasets. Missing values in the dataset were treated as noise, and the DAE reconstructed the missing values. The authors first replaced the missing entries in the dataset with 0s. Then, the DAE was trained using the loss function for categorical and numerical variables to minimize the reconstruction errors. Finally, the imputed dataset was obtained by predicting missing values (which were initially replaced with 0). The authors used simulated data, adult datasets [49], and a Cooperative Congressional Election Study dataset [50]. The authors showed that the proposed method achieves better efficiency and accuracy compared to other works. Similar work was done by Gondara and Wang [12], who proposed the DAE for missing value predictions using 15 publicly available datasets. The authors compared the method with the MICE imputation and showed that the DAE was better.

Deep learning-based GAN networks have also achieved attention for missing data imputation such as the work by Yoon et al., who proposed Generative Adversarial Imputation Nets (GAIN). This uses the GAN architecture with an additional hint to the generator to ensure the generator creates the new synthetic data based on the original input data distribution. The authors showed that the proposed GAIN achieves better performance compared to other imputation methods using four datasets. Although GAIN achieved promising performance, it ignores category information, which identifies the relationship between samples. To address this, Wang et al. proposed the pseudo-label conditional GAIN (PC-GAIN) [51] to consider category information and improve the imputation. The authors first used the low-missing-rate samples to impute using GAN followed by clustering to obtain the pseudo-labels. The pre-trained classifier was obtained from the imputed dataset and corresponding pseudo-labels. Finally, the full dataset was imputed using GAN, which used a pre-trained classifier for the generator. The results showed that the proposed method achieves better performance compared to other imputation techniques. Work done by Awan et al. [42] proposed conditional GAIN (CGAIN) which considers the class-specific distribution for missing data imputation, especially in imbalanced datasets. The authors show that their proposed work achieved better performance compared to other algorithms. Although GAN-based methods are used for missing data imputation, no work has been done to study the effects of the training data on imputation performance. Therefore, we propose a GAN-based imputation that increases the amount of training data by adding synthetic samples to improve missing value predictions in the testing data. The details of the proposed methodology are described below.

### III. PROPOSED METHODOLOGY

This section presents the proposed methodology. We first describe the formulation of our method for mixed data imputation. We then propose an algorithm to handle the
challenge of imputing missing values in a mixed dataset. Then, we explain the algorithm by first describing the oversampling of training data using GANs. Finally, we briefly explain the imputation techniques for mixed datasets.

Imputing missing values is important to reduce bias due to missingness [52]. Consider the dataset $X = \{X_{obs}, X_{miss}\}$, where $X_{obs}$ contains entries without missing values and $X_{miss}$ is data with missing entries. As most of ML-based algorithms require complete observations [53], one way to handle $X_{miss}$ is to completely delete it, but this will result in information loss and a lower efficiency compared to imputation [54]. Thus, the objective is to impute missing values in $X_{miss}$ such that they closely resemble the complete data. Traditionally, missing values can be imputed using traditional statistical imputation methods such as the mean, median, or mode; however, these usually result in lower accuracies [15]. Therefore, more sophisticated techniques such as ML- and DL-based algorithms are used to learn from the observed data $X_{obs}$ and impute the missing $X_{miss}$ by predicting the missing values [14], [15], [55], [56].

Moreover, many algorithms are available to impute numerical values. However, mixed (numerical and categorical) data imputation is challenging due to the complex joint data distribution [53]. In real applications, datasets are mixed and contain missing values due to various reasons. Therefore, the objective of this study is to impute mixed datasets using model-based approaches. The proposed methodology is represented in algorithm 1. First, the dataset $X = \{X_{obs}, X_{miss}\}$ is divided into observed values $X_{obs}$ and missing values $X_{miss}$. As model-based imputation algorithms require sufficient training data for better generalization [57], we aim to increase the amount of training data $X_{obs}$ using state-of-art GANs. After training the GAN with $X_{obs}$, we achieved synthetic samples of $X_{GAN}$ as shown in 4-5 of algorithm 1. The GANs focus primarily on unstructured data, such as generating images [58], [59], [60], but most real-world datasets are tabular [61]. Thus, obtaining mixed synthetic data is more challenging because GANs need to generate categorical and numerical columns simultaneously by applying the $softmax$ and $tanh$ functions at the output layer. Therefore, there are a limited number of GANs that are capable of generating mixed tabular data [61]. We used two state-of-the-art GANs to synthesize mixed training data. The $X_{obs}$ and $X_{GAN}$ are combined to obtain a lagar dataset, which will be used to train the imputation model $M$. Finally, the missing values in $X_{miss}$ are predicted using $M$, and the imputed dataset $X_{imputed}$ is obtained. The proposed methodology is also represented in Figure 1.

Algorithm 1 Mixed Data Missing Data Imputation Using GANs

1. The input $X$ contains observed values $X_{obs}$ and missing values $X_{miss}$
2. The output $X_{imputed}$ dataset with missing entries is predicted
3. Divide $X$ into $X_{obs}$ and $X_{miss}$
4. Use the GANs to generate $X_{GAN}$ from $X_{obs}$
5. Use $X_{obs}$ to train the GANs to generate synthetic samples: $X_{GAN}$
6. Combine $X_{GAN}$ and $X_{obs}$ to obtain training data $X_{Train}$
7. Train the imputation algorithms on $X_{Train}$ to obtain imputation model $M$
8. Impute $X_{miss}$ using $M$
9. $X_{imputed}$ with imputed values

A. GANS FOR TABULAR DATA

We adopted two different versions of the GANs (tabular [40] and conditional [41]) for mixed data synthesis. Each GAN is briefly described below.

1) TABULAR GAN

Generating tabular data is a challenge compared to images as they contain various types of data, such as categorical, numerical, text, time, and cross-table references [40]. To generate synthetic data using TGAN, consider a table $T$ with mixed data (numerical $\{n1, n2, \ldots\}$ and categorical $\{c1, c2, \ldots\}$) that have a joint probability distribution of $P(n1, n2 \ldots C1, c2 \ldots)$ with each row being one sample distribution from the joint distribution [40]. Both the numerical and categorical data are preprocessed such that the numerical values in the dataset are normalized to $[-1, 1]$. Furthermore, to handle multimodal distributions, mode-specific normalization with five modes is used based on the Gaussian mixture model (GMM) [40]. For categorical data generation, a sample from $|C|\rightarrow$ as the categorical variable is first converted into a one-hot vector $c_{ij}$, and uniform noise is added to each dimension followed by renormalizing the representation as $C_{imputed} = c_{ij} / \sum_{k=1}^{C} d_{ij,k}$.

After preprocessing, the sum of both vectors for the categorical and numerical columns is sent to the discriminator and generator as the input and output, respectively. The generator in the TGAN is a long-short term memory (LSTM) network while a multi-layered perceptron (MLP) is used in the discriminator. One of the disadvantages of the TGAN is that it requires a relatively long time to train [31], [62].
2) CONDITIONAL TABULAR GAN

The CTGAN is based on PacGAN [63] and uses a conditional generator and training-by-sampling approach for synthetic tabular data. Unlike TGAN, which uses GMM, the authors adopt a variational mixture model (3 CTGAN variants) to automatically estimate the number of modes for the numerical variables. Furthermore, a conditional generator is introduced to handle the category imbalance in the categorical variables. Thus, the categories are sampled evenly for the variables during the training process. The CTGAN introduces three key elements: conditional vector, generator loss, and training-by-sampling method. Firstly, a conditional vector is introduced to select an entry from a categorical column. For example: let $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5, 6\}$ be two categorical columns. The condition $C_2 = 2$ gives the conditional vector of $\text{cond} = [0, 0, 0, 1, 0]$. After the conditional vector, the generator produces any categorical one-hot vector regardless of the conditional vector defined as $C_2 = 2$; therefore, the conditional generator is enforced by penalizing its loss using the Wasserstein GAN loss with a gradient penalty [64] so it learns to generate an exact copy as specified by the conditional vector. Finally, in the training-by-sampling step, the conditional generator is assessed by the critic to estimate the distance between the conditional distribution on the real data and the learned conditional distribution.

B. DATA IMPUTATION ALGORITHMS

To effectively evaluate our proposed methodology, we select three missing imputation algorithms as a baseline: MICE [65], MissForest [66], and DAEs [15]. We propose that using GANs with the incorporated training data improves the imputation performance of the baseline algorithms. The baseline methods are briefly described below.

1) MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE)

The MICE algorithm uses multiple imputations to provide missing values in mixed datasets using four general steps [65]. First, the missing values in the datasets are imputed with initial values as a temporary placeholder, which includes the mean. Second, the placeholder for one variable is set to missing while the placeholders for other variables remain filled. The missing values in the second step are regressed via linear regression and are dependent variables. Fourth, the fitted regression model is used to predict the missing values of the other variables. These four steps are used interactively to predict the missing values using a regression model. The details about the MICE algorithm can be obtained from [65].

2) MISS FOREST

The MissForest is based on the random forest algorithm for mixed data imputation and first imputes the missing values with mean imputation as an initial guess value. The variables are then first sorted based on minimum missing values. The missing values are predicted by training the RF model with the observed data. The imputation procedure is repeated iteratively until the stopping criterion is met [66].

3) DENOISING AUTO-ENCODERS (DAE)

Autoencoders are neural networks that are used to reproduce a copy of the input from the encoding. The input is first encoded into the lower-dimensional latent representation. Then, the latent representation is back-decoded at the output with the same dimension and distribution as the input. To avoid the identical representation of data, one variant of the autoencoder is the DAEs where the input $X$ is corrupted by the injecting noise. The corrupted input $X_{\text{noise}}$ is then encoded, and the decoder reproduces the denoised version $X_{\text{denoise}}$ of the encoded input. The noise is reconstructed and the $X_{\text{denoise}}$ is similar to $X$. For our missing data problem, we treat the missing values in $X_{\text{miss}}$ as noise and aim to reconstruct the missing values. We follow the method proposed by Lall et al. [15] to reconstruct the missing values $X_{\text{miss}}$.

IV. EXPERIMENTS AND RESULTS

The experimental setup is represented in Figure 2. We conduct a series of experiments on four publicly available mixed datasets from different domains of social, financial, and health. Each dataset is divided into multiple ratios of training $X_{\text{Train}}$ and testing $X_{\text{Test}}$ sets. In the testing set $X_{\text{Test}}$, we introduce missing data with different missingness ratios and obtain $X_{\text{miss}}$. The training data act as $X_{\text{Obs}}$, which is oversampled using TGAN and CTGAN. Finally, the imputation algorithms are trained with the $X_{\text{Train}}$ and the missing values in $X_{\text{miss}}$ are predicted to obtain $X_{\text{Imputed}}$. The reconstruction errors are calculated using the RMSE for each categorical and numerical variable.
TABLE 2. Datasets used in this study.

| Dataset            | Domain       | Records | Total Features | Numerical features | Categorical features |
|--------------------|--------------|---------|----------------|--------------------|----------------------|
| Adult              | Social       | 48,842  | 14             | 6                  | 8                    |
| Heart              | Health       | 363     | 15             | 7                  | 6                    |
| German Credit      | Financial    | 1000    | 20             | 13                 | 7                    |
| Australian Credit  | Financial    | 690     | 14             | 8                  | 6                    |

A. DATASET

Four different real-world datasets from domains such as social, financial, and health are used to evaluate the performance of the proposed method. The datasets are publicly available and adopted from the UCI directory [30]. The statistics of the datasets are shown in Table 2.

1) ADULT

The adult dataset is mixed with categorical and numerical variables obtained from the 1994 United States Census with 48,842 individuals from 42 countries having 14 characteristics. The prediction task using the adult dataset is to determine if a person makes more than $50 thousand a year. The dataset contains 14 features that consist of 6 numerical and 8 categorical.

2) GERMAN CREDIT

The German credit dataset classifies whether a person has good or bad credit risks based on 20 mixed attributes related to banking and financial details. The dataset includes 13 categorical and 7 numerical variables. There are 1000 instances that belong to two classes in the dataset, i.e., good credit means the customer is likely to repay the loan, while bad credit is for the customers who are not. Not approving the load to good customers may result in a loss of business while approving a loan to bad customers results in financial losses for the bank.

3) AUSTRALIAN CREDIT

The Australian credit dataset provides information related to credit card approval based on 14 attributes. There is a total of 690 credit card applications (instances) with 14 attributes consisting of 8 categorical and 6 numerical.

4) HEART CLEVELAND

Heart Cleveland is a small medical domain dataset that contains 302 instances. There are 13 different attributes with 7 categorical and 6 numerical.

B. TRAINING AND TESTING DIVISION

To illustrate the effect of the proposed method on the considered data, each dataset is randomly divided into different training and testing sets. Multiple training-testing divisions include 80-20, 60-40, 40-60, and 20-80. Five different sets were obtained for each division to generalize the proposed method.

C. MISSINGNESS

The datasets contain a small amount of missingness. Therefore, we inject artificial missingness using the MCAR mechanism used by [14], [20], [47], [67], [51]. Missing data imputation in each testing set is evaluated by varying the missing rates at 10%, 30%, 50%, and 70%.

D. GANS FOR DATA SYNTHESIS

We train CTGAN and TGAN using $X_{obs}$ to generate synthetic samples $X_{GAN}$. To determine the effect of the oversampling ratio on the training data, different $X_{GAN}$ ratios are added with $X_{obs}$. We first add 25% of the samples from $X_{GAN}$ into $X_{obs}$ and repeat the experiment for 50% and 75% ratios. Finally, we doubled the training data that contain an equal number of $X_{GAN}$ and $X_{obs}$ samples. If we increase $X_{obs}$ by 25% using TGAN, the experiment is denoted as TGAN 25%.

E. PERFORMANCE METRICS

The root means squared error (RMSE) is a performance metric to evaluate the actual and predicted values using the proposed imputation model. The RMSE for both numerical ($RMSE_{num}$) and categorical ($RMSE_{cat}$) values are calculated separately, as represented in Eqs. (1) and (2), respectively.

\[
RMSE_{num} = \sqrt{\frac{1}{|N_{num}|} \sum_{(i,j) \in N_{num}} (\hat{x}_{ij} - x_{ij})^2}
\]

\[
RMSE_{cat} = \frac{1}{|N_{cat}|} \sum_{(i,j) \in N_{cat}} 1(|i,j| \neq x_{ij})
\]

F. RESULTS

To evaluate the performance of the proposed work, we considered our results on four datasets for multiple training-testing divisions with various missing ratios. Each experiment was repeated five times at every step on the shuffled dataset, and the averaged results are represented in Tables 3–16. Only the best results are represented in this paper, while the detailed experimental results are provided in the supplementary data Table S1–S50.

The experiments performed on the German dataset with the 80-20 training-testing set are represented in Table S1–S12. For the MICE imputation, the best (smallest) RMSE for the numerical variables was achieved by CTGAN with 25% oversampled data for a 10% missingness. For the categorical variables, the best RMSE was achieved by TGAN with 50% oversampled data followed by CTGAN with 25%. Similarly, TGAN achieved a better performance in the MissForest categorical while the baseline for the numerical was best in the MissForest imputation. TGAN with a 100% oversampling ratio achieved the best performance using the DAE imputation in the numerical variable. The performance using TGAN 25% for categorical variables was the best while TGAN with
Table 3. Imputation results for the 70% missingness in the 80-20 data division.

| German Dataset | MICE Imputation | MissForest | DAEs |
|----------------|----------------|------------|------|
|                | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 498.415 | 0.4091 | 487.117 | 0.3358 | 442.604 | 0.3786 |
| TGAN 25%       | **491.146** | **0.4060** | 533.916 | **0.3557** | **452.758** | **0.3729** |
| TGAN 50%       | 493.284 | 0.4101 | 516.056 | 0.3598 | 439.273 | 0.3805 |
| TGAN 75%       | 568.581 | 0.4117 | 518.024 | 0.3349 | 426.445 | 0.3772 |
| TGAN 100%      | 551.549 | 0.4092 | 540.120 | 0.3364 | 419.891 | 0.3795 |
| CTGAN 25%      | 484.878 | 0.4099 | 493.359 | 0.3586 | 456.493 | 0.3914 |
| CTGAN 50%      | 507.172 | 0.4099 | 523.997 | 0.3684 | 456.493 | 0.3914 |
| CTGAN 75%      | 532.211 | 0.4194 | **474.360** | 0.3543 | 437.900 | 0.4145 |
| CTGAN 100%     | 511.707 | 0.4259 | 564.009 | 0.3553 | 445.770 | 0.4138 |

Table 4. Imputation results for the 70% missingness in the 60-40 data division.

| German Dataset | MICE Imputation | MissForest | DAEs |
|----------------|----------------|------------|------|
|                | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 517.419 | 0.4092 | 492.582 | 0.3711 | 449.657 | 0.3865 |
| TGAN 25%       | **519.356** | **0.4082** | 486.916 | 0.3584 | 449.968 | 0.3825 |
| TGAN 50%       | 516.732 | 0.4090 | 485.008 | 0.3641 | 438.829 | 0.3860 |
| TGAN 75%       | 523.082 | 0.4118 | **477.392** | **0.3409** | 434.766 | 0.3843 |
| TGAN 100%      | 528.647 | 0.4106 | 492.702 | 0.3412 | **425.456** | **0.3810** |
| CTGAN 25%      | 495.660 | 0.4099 | 482.874 | 0.3620 | 448.034 | 0.3950 |
| CTGAN 50%      | **488.753** | **0.4112** | 490.159 | 0.3770 | 441.635 | 0.3858 |
| CTGAN 75%      | 509.008 | 0.4157 | 497.808 | 0.3592 | 442.746 | 0.4142 |
| CTGAN 100%     | 501.009 | 0.4185 | 467.616 | 0.3461 | 432.705 | 0.4021 |

100% also achieved a comparable performance. For the 30% missingness represented in Table S2, the performance of the numerical variable baseline for MICE and MissForest was the best, while CTGAN was better for categorical variables for MICE and TGAN for MissForest. For imputation using DAE, both RMSE (numerical and categorical) were best with TGAN. For a 50% missingness, CTGAN performed better for MICE and MissForest under both categorical variables, except MissForest in which TGAN performed better. For imputation using DAE, the TGAN achieved a better performance in both categorical and numerical variables with an RMSE of 0.316 and 363.077, respectively. Similarly, in the 70% missingness for the 80-20 training-testing represented in Table 3, the performance of the proposed method was better compared to the baseline methods, in two GANs with imputation using MissForest, CTGAN achieved the smallest RMSE for the numerical variables. In the remaining imputations, TGAN was better. Overall, the GAN approach is better compared with the baseline methods.

The results obtained from the German dataset on the 60-40 training-testing data are shown in Table 4. For all missingness ratios, the performance of the proposed method is better than the baseline methods. Only when there was 30% missingness was the numerical RMSE of the MissForest slightly better than the GAN-based method. The results represented in Table 4 show that TGAN achieved the best performance compared to all baseline methods. The CTGAN was only better for numerical imputation using the MICE imputation while TGAN achieved the best performance overall. For the 50% missingness (Table S 6), CTGAN was better for MICE imputation, and CTGAN was better for numerical data imputation using DAE.

For the 20-80 data division, the results for the 70% missingness are shown in Table 6. For the MICE, the CTGAN performed well in both numerical and categorical imputation. For DAE and MissForest, the TGAN achieved a lower RMSE for both categorical and numerical imputation, while the MissForest baseline achieved a better numerical RMSE. Similar performances were achieved at 50% where only the numerical imputation using MissForest was better. For 10% and 30%, the numerical imputation using the baseline method was better for MICE and MissForest imputations (see Tables S10–S12).

The experiments performed on the Australian credit dataset are shown below, which illustrates that the proposed method improves performance compared to the baseline. For the 80-20 training-testing division, Table 7 shows that TGAN achieved the best overall results for the 50% missingness compared to all baseline methods. For imputation using DAE, the CTGAN performed slightly better than TGAN with a numerical RMSE of 473.41 compared to 474.07. For MICE, imputation, the baseline achieved 566.69 and
TABLE 5. Imputation results for the 70% missingness in the 40-60 data division.

| German Dataset | MICE Imputation | MissForest | DAEs |
|----------------|-----------------|------------|------|
|                | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 497.368         | 0.4102     | 488.140  | 0.3713     | 453.523 | 0.3847 |
| TGAN 25%       | 494.137         | 0.4117     | 494.428  | 0.3584     | 448.372 | 0.3824 |
| TGAN 50%       | 493.011         | 0.4122     | 484.414  | 0.3620     | 448.266 | 0.3779 |
| TGAN 75%       | 484.569         | 0.4128     | 493.372  | 0.3426     | 442.728 | 0.3754 |
| TGAN 100%      | 510.918         | 0.4128     | 502.865  | **0.3419** | **437.162** | 0.3882 |
| CTGAN 25%      | 499.198         | **0.4102** | 510.196  | 0.3607     | 449.317 | 0.3954 |
| CTGAN 50%      | 491.312         | 0.4117     | 494.591  | 0.3703     | 451.620 | 0.3975 |
| CTGAN 75%      | 495.905         | 0.4150     | 490.797  | 0.3536     | 448.760 | 0.4097 |
| CTGAN 100%     | 501.306         | 0.4153     | 488.699  | 0.3799     | 447.122 | 0.4021 |

TABLE 6. Imputation results for the 70% missingness in the 20-80 data division.

| German Dataset | MICE Imputation | MissForest | DAEs |
|----------------|-----------------|------------|------|
|                | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 479.438         | 0.4122     | 456.741  | 0.3666     | 457.513 | 0.3832 |
| TGAN 25%       | 484.893         | 0.4125     | 496.955  | 0.3569     | 451.725 | 0.3815 |
| TGAN 50%       | 483.952         | 0.4146     | 522.572  | 0.3638     | **444.993** | **0.3742** |
| TGAN 75%       | 497.848         | 0.4139     | 496.501  | 0.3404     | 445.761 | 0.3821 |
| TGAN 100%      | 479.189         | 0.4157     | 490.615  | **0.3394** | 452.020 | 0.3759 |
| CTGAN 25%      | 467.758         | 0.4086     | 497.411  | 0.3618     | 458.197 | 0.4003 |
| CTGAN 50%      | 482.666         | 0.4092     | 501.747  | 0.3691     | 452.931 | 0.3992 |
| CTGAN 75%      | 482.546         | **0.4081** | 501.827  | 0.3499     | 457.142 | 0.4066 |
| CTGAN 100%     | 484.140         | 0.4104     | 480.349  | 0.3490     | 459.825 | 0.4055 |

TABLE 7. Imputation results for the 50% missingness in the 80-20 data division.

| Australian Dataset | MICE Imputation | MissForest | DAEs |
|--------------------|-----------------|------------|------|
|                    | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method    | 566.698         | 0.3172     | 498.842  | 0.2755     | 476.284 | 0.2918 |
| TGAN 25%           | 543.578         | 0.3194     | 472.433  | 0.3040     | 476.695 | 0.2952 |
| TGAN 50%           | 508.814         | 0.3193     | **393.484** | 0.3124     | 475.622 | 0.3013 |
| TGAN 75%           | 546.887         | 0.3177     | 407.795  | **0.2754** | 474.075 | **0.2906** |
| TGAN 100%          | **496.621**     | **0.3137** | 467.376  | 0.2817     | 475.629 | 0.2963 |
| CTGAN 25%          | 1311.135        | 0.3263     | 470.276  | 0.3231     | 475.241 | 0.3200 |
| CTGAN 50%          | 964.241         | 0.3285     | 398.377  | 0.3341     | 476.362 | 0.3185 |
| CTGAN 75%          | 1210.111        | 0.3215     | 573.823  | 0.2979     | **473.416** | **0.3201** |
| CTGAN 100%         | 1259.792        | 0.3281     | 605.661  | 0.3042     | 474.260 | 0.3252 |

0.317 for the numerical and categorical variables, respectively, while TGAN achieved 496.62 and 0.313. Similarly, for MissForest, the baseline method achieved 498.84 and 0.275 for the numerical and categorical variables, respectively, while TGAN achieved 393.48 and 0.27. Finally, for imputation using DAE, the CTGAN achieved an RMSE of 473.41 compared to 476.28 by the baseline method, while the categorical RMSE was 0.290 by TGAN compared to 0.291 by the baseline method. The results for 10%, 30%, and 70% missingness ratios represented in Tables S13, S14, and S15, respectively, show that incorporating GANS with the baseline method improves the imputations.

The results obtained from the 60-40 data divisions using the Australian credit data show that for the 70% missingness, the proposed method achieved a better imputation compared to the baseline. Moreover, for missingness of 10-50%, the proposed method had a relatively better imputation of missing values. Table 8 represents that for the 70% missingness, the TGAN with an oversampling ratio of 75% achieved the best performance compared to baseline and CTGAN. The results in Table S16–S18 show that for the 10-30%, the performances of the CTGAN and TGAN were better than the baseline methods. For the 50% missingness, the performance of TGAN was better than the baseline methods (except for the categorical
The experiments performed on the heart dataset show that the use of GAN improves the overall predictions of missing values. For the 80-20 data division, the 10-50% missingness represented in Tables S27–S29 shows that the overall imputation with the aid of TGAN improves the prediction of missing values. Thus, GAN-based imputation performs either better or is comparable to the baseline methods. The best results obtained for the 70% missingness are shown in Table 9, where the TGAN achieved the best performance compared to CTGAN and baseline methods. Only the MissForest baseline method was slightly better with an RMSE of 0.452 compared to 0.453 with TGAN. Better performances were achieved when the training data increased with 100% TGAN-generated samples. Similar performance was achieved for the 60-40 data division. For the 70%, missingness shown in Table 10, incorporating synthetic TGAN-generated data improved the imputing missing values compared to the baseline method. Only the categorical RMSE was better than the proposed method using 70% missingness. For the 60-40 data division, the 10-50% missingness ratios also show that the proposed method achieved better performance, as depicted in Tables S30–S32.

For the 40-60 data division, the results obtained from the 10% missingness are shown in Table 11, where using the proposed method helps better reconstruct missing values. The proposed method achieved better results for both categorical and numerical values compared to the baseline methods. The results depicted in Tables S33–S35 show that the proposed method achieved better or similar performances compared to the baseline methods. The results obtained from the 20-80 data division show similar performances where the GAN improves the prediction performance compared to the baseline methods. For instance, for 10% missingness results shown in Table 12 for DAE imputation have baseline categorical and numerical RMSEs of 0.166 and 16.33, respectively, while the TGAN achieved 0.158 and 14.57. Similarly, the results for 30-70% missingness ratios in Tables S36–S38 suggest that incorporating GANS improves the reconstruction error.

The experiments performed on the large adult dataset, which contains over 48,000 samples. The best results obtained on the adult dataset for the 80-20 data division using the 70% missingness are represented in Table 13, which shows that CTGAN achieved the best imputation performance when used with the MissForest imputation. However, for categorical imputation, the CTGAN performed better. The results shown in Tables S39–S41 for 10-50% missingness also indicate that using GAN for imputation achieves a lower RMSE compared to the baseline. Similar performances were achieved for the 60-40 data division in which the MissForest imputation baseline method was improved with CTGAN, the MICE baseline was improved using TGAN for RMSE using MissForest, which was slightly better). Results obtained from the 40-60 and 20-80 data divisions for the Australian credit dataset are represented in Tables S19–S26, which show that the proposed method did not perform as well.

**TABLE 8.** Imputation results for the 70% missingness in the 60-40 data division.

| Australian Dataset | MICE Imputation | MissForest | DAEs |
|--------------------|----------------|------------|------|
|                    | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method    | 895.670    | 0.3974     | 993.388    | 0.3394     | 790.325    | 0.3440     |
| TGAN 25%           | 836.683    | 0.3966     | 788.421    | 0.3581     | 788.717    | 0.3484     |
| TGAN 50%           | 821.430    | 0.4012     | 750.250    | 0.3620     | 789.188    | 0.3524     |
| TGAN 75%           | 834.819    | 0.3900     | 677.169    | 0.3405     | 788.424    | 0.3433     |
| TGAN 100%          | 853.416    | 0.3953     | 838.240    | 0.3428     | 789.083    | 0.3475     |
| CTTGAN 25%         | 1123.382   | 0.4008     | 803.895    | 0.3622     | 788.840    | 0.3693     |
| CTTGAN 50%         | 1333.533   | 0.4056     | 682.660    | 0.3819     | 788.482    | 0.3679     |
| CTTGAN 75%         | 1489.756   | 0.4037     | 687.046    | 0.3616     | 788.526    | 0.3764     |
| CTTGAN 100%        | 1445.027   | 0.4062     | 801.372    | 0.3607     | 788.484    | 0.3788     |

**TABLE 9.** Imputation results for the 70% missingness in the 80-20 data division.

| Heart Dataset | MICE Imputation | MissForest | DAEs |
|---------------|----------------|------------|------|
|               | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 16.902     | 0.4735     | 18.037     | **0.4521**  | 69.962     | 0.4434     |
| TGAN 25%      | 17.233     | **0.4656** | 18.859     | 0.4718      | 67.787     | 0.4534     |
| TGAN 50%      | 17.845     | 0.4708     | 18.355     | 0.4752      | 67.073     | 0.4466     |
| TGAN 75%      | 16.869     | 0.4795     | 18.506     | 0.4552      | 66.652     | 0.4451     |
| TGAN 100%     | **16.737** | 0.4764     | **17.836** | 0.4537      | **64.862** | **0.4397** |
| CTTGAN 25%    | 18.838     | 0.4700     | 19.428     | 0.4697      | 68.157     | 0.4565     |
| CTTGAN 50%    | 20.093     | 0.4858     | 21.025     | 0.4858      | 67.678     | 0.4497     |
| CTTGAN 75%    | 21.421     | 0.4948     | 20.380     | 0.4860      | 67.359     | 0.4476     |
| CTTGAN 100%   | 20.149     | 0.4885     | 21.655     | 0.4917      | 65.913     | 0.4514     |
TABLE 10. Imputation results for the 70% missingness in the 60-40 data division.

| Heart Dataset | MICE Imputation | MissForest | DAEs |
|---------------|-----------------|------------|------|
|               | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 17.062 0.4671 | 18.062 0.4457 | 67.728 0.4465 |
| TGAN 25%      | 16.387 0.4743 | 18.534 0.4645 | 66.386 0.4443 |
| TGAN 50%      | 17.643 0.4737 | 17.876 0.4643 | 65.823 0.4459 |
| TGAN 75%      | 17.584 0.4770 | 17.645 0.4410 | 64.648 0.4434 |
| TGAN 100%     | 17.014 0.4800 | 18.610 0.4401 | **63.938 0.4404** |
| CTGAN 25%     | 17.617 0.4811 | 18.238 0.4843 | 66.641 0.4454 |
| CTGAN 50%     | 18.597 0.4889 | 18.953 0.4946 | 66.490 0.4522 |
| CTGAN 75%     | 18.751 0.4793 | 19.496 0.4856 | 65.444 0.4535 |
| CTGAN 100%    | 19.706 0.4966 | 19.415 0.4861 | 64.840 0.4547 |

TABLE 11. Imputation results for the 10% missingness in the 40-60 data division.

| Heart Dataset | MICE Imputation | MissForest | DAEs |
|---------------|-----------------|------------|------|
|               | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 5.957 0.1695 | 6.571 0.1635 | 16.470 0.1694 |
| TGAN 25%      | 6.112 0.1723 | 6.788 0.2479 | 15.985 0.1686 |
| TGAN 50%      | 6.398 0.1697 | 6.819 0.2375 | 16.267 **0.1646** |
| TGAN 75%      | 6.154 0.1758 | 6.901 0.1634 | 14.789 0.1667 |
| TGAN 100%     | **5.950 0.1772** | 7.074 **0.1619** | **14.102 0.1658** |
| CTGAN 25%     | 6.132 0.1731 | 6.836 0.2517 | 16.043 0.1671 |
| CTGAN 50%     | 6.224 0.1681 | **6.443 0.2774** | 16.387 0.1675 |
| CTGAN 75%     | 6.237 **0.1680** | 6.751 0.1728 | 14.966 0.1700 |
| CTGAN 100%    | 6.524 0.1737 | 6.891 0.1741 | 14.383 0.1750 |

TABLE 12. Imputation results for the 10% missingness in the 20-80 data division.

| Heart Dataset | MICE Imputation | MissForest | DAEs |
|---------------|-----------------|------------|------|
|               | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 6.002 0.1658 | 7.059 0.1595 | 16.336 0.16606 |
| TGAN 25%      | 5.976 **0.1650** | 7.182 0.2442 | 15.622 0.1622 |
| TGAN 50%      | 6.096 0.1706 | 6.929 0.2624 | 15.631 **0.1582** |
| TGAN 75%      | 6.076 0.1741 | 6.839 **0.1552** | 15.593 0.1605 |
| TGAN 100%     | **5.743 0.1688** | **6.836 0.1563** | **14.576 0.1625** |
| CTGAN 25%     | 6.325 0.1695 | 7.216 0.2441 | 15.688 0.1622 |
| CTGAN 50%     | 5.926 0.1730 | 7.662 0.2721 | 15.768 0.1631 |
| CTGAN 75%     | 6.664 0.1685 | 8.063 0.1690 | 15.794 0.1613 |
| CTGAN 100%    | 6.863 0.1725 | 7.687 0.1705 | 14.743 0.1662 |

Numerical and CTGAN for categorical imputation, and the TGAN aided categorical imputation when DAE was used as a baseline. The results for the 70% missingness are shown in Table 14 while results for 10-50% missingness are shown in Tables S42–S44, which indicates the proposed method improves the imputation performance compared to the baseline.

The results for the 10% missingness in the 40-60 data division for the adult dataset are represented in Table 15, where oversampling using GANs achieved better performance compared to the baseline methods. The results show that CTGAN achieved a better numerical RMSE for the MissForest and categorical for MICE, while the TGAN achieved a better RMSE for categorical variables in the MissForest and DAE imputation. The numerical imputation was better when using the MICE algorithm. Similarly, for 30-70% missingness, the proposed method achieved relatively better performance compared to the baseline, as shown in Tables S45–S47. Finally, for the 20-80 data division, the 50% missingness shown in Table 16 indicates that the TGAN achieved the best numerical imputation results compared to the baseline and CTGAN. For categorical imputation, the TGAN was only better for DAEs while the CTGAN was better for MICE imputation; the RMSEs were similar to the baseline in the MissForest imputation. Similarly, other missing ratios, such as 10, 30, and 70% represented in Tables S48–S50, show
that the proposed method achieved a better imputation performance compared to the baseline methods.

V. DISCUSSION

In real-life scenarios, most datasets consist of mixed data with frequent missing entries. In addition, small datasets are usually publicly available; therefore, training an ML model with a limited amount of data does not achieve promising results. Moreover, missing value imputation in mixed datasets is challenging and has achieved little attention. This work proposes a method to improve the prediction performance of missing values when training with oversampled data. Oversampled training data is attained using two types of GANs for mixed data synthesis: TGAN and CTGAN. We have shown that using various experiments our proposed method works on various mixed datasets using different imputation algorithms.

We used four mixed datasets that are publicly available in the UCI directory [30]. Each dataset was divided into training-testing sets {80-20, 60-40, 40-60, and 20-80}. In each test set, missingness was introduced with a different missing ratio of 10–70%. Each experiment was performed five times and the average results are presented here. For the German credit dataset, the overall best performance was achieved when the missingness was 70% (see Tables 3–6) followed by 50% (see Tables S3, S6, S9). For 10–30% missingness, the proposed method performed better than or similar to the baseline methods. Similar imputation results were achieved when the proposed method was applied to the Australian credit dataset with a 70% missingness (see Table 8) with the proposed method performing best. For 20-80 and 60-40 data divisions, the proposed method performed better or similar to the baseline methods for 10–50% missingness ratios. However, for 40-60 and 20-80 data divisions, our algorithm did not perform as well. For the heart dataset, the proposed method achieved the best performance for all missingness ratios relative to the baseline methods (Tables 9–12 and S27–S38). Finally, for the adult dataset, our proposed method outperformed the baseline methods in all data divisions with nearly all the missingness ratios.

We found through a series of experiments that our method was better than the baseline methods at nearly all missingness ratios. However, the proposed method works better for a 70% missingness, which suggests that the method can be applied when the data contains significant missing entries. We also compared each baseline method between themselves as we used default parameters without any hyperparameter optimization, and the presented results reflect this comparison. For the German credit dataset, the numerical imputation using the baseline MICE was better when the missingness was 10–30%. When the missingness was more than 50%, then baseline DAE performed well when reconstructing the numerical values. Categorical imputation using the MissForest was better than the MICE and DAE imputation. For the
TABLE 15. Imputation results for the 10% missingness in the 40-60 data division.

| Adult Dataset | MICE Imputation | MissForest | DAEs |
|---------------|-----------------|------------|------|
|               | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 8217.131        | 0.0707     | 6571.992   | 0.0766     | 6343.134   | 0.0734     |
| TGAN 25%      | 8217.867        | 0.0709     | 9039.662   | 0.0765     | 6346.530   | 0.0742     |
| TGAN 50%      | 8151.255        | 0.0713     | 9746.813   | 0.0767     | 6385.015   | 0.0733     |
| TGAN 75%      | 8367.069        | 0.0717     | 9740.203   | 0.0765     | 6449.737   | 0.0748     |
| TGAN 100%     | 8379.925        | 0.0722     | 9054.330   | 0.0763     | 6476.684   | 0.0727     |
| CTGAN 25%     | 11945.384       | 0.1021     | 6542.574   | 0.0770     | 6362.178   | 0.0753     |
| CTGAN 50%     | 8216.849        | 0.0685     | 6527.302   | 0.0772     | 6359.393   | 0.0731     |
| CTGAN 75%     | 8221.778        | 0.0681     | 6548.693   | 0.0772     | 6346.646   | 0.0748     |
| CTGAN 100%    | 8189.125        | 0.0677     | 6577.652   | 0.0772     | 6332.608   | 0.0755     |

TABLE 16. Imputation results for the 50% missingness in the 20-80 data division.

| Adult Dataset | MICE Imputation | MissForest | DAEs |
|---------------|-----------------|------------|------|
|               | RMSE (Num)      | RMSE (Cat) | RMSE (Num) | RMSE (Cat) | RMSE (Num) | RMSE (Cat) |
| Baseline Method | 18501.733       | 0.2771     | 15036.281 | 0.1916     | 14238.286  | 0.2534     |
| TGAN 25%      | 18422.686       | 0.2774     | 14093.102 | 0.1918     | 14236.192  | 0.2513     |
| TGAN 50%      | 18210.075       | 0.2778     | 13949.401 | 0.1918     | 14228.827  | 0.2521     |
| TGAN 75%      | 18080.549       | 0.2783     | 13946.616 | 0.1919     | 14197.490  | 0.2507     |
| TGAN 100%     | 18089.489       | 0.2785     | 14058.192 | 0.1917     | 14279.846  | 0.2529     |
| CTGAN 25%     | 18820.926       | 0.2731     | 14926.069 | 0.1918     | 14371.149  | 0.2568     |
| CTGAN 50%     | 18739.993       | 0.2988     | 14983.610 | 0.1916     | 14365.937  | 0.2539     |
| CTGAN 75%     | 18652.984       | 0.2986     | 14192.970 | 0.1923     | 14385.511  | 0.2563     |
| CTGAN 100%    | 18877.348       | 0.2984     | 14870.632 | 0.1976     | 14412.648  | 0.2545     |

Australian credit dataset, categorical imputation using MissForest was better overall missingness ratios and data divisions and achieved better performance for numerical imputation when the data divisions were 80-20 and 60-40. The performance using DAE was better when the data divisions were 40-60 and 20-80.

For the heart dataset, the baseline MICE was best for numerical imputation. For categorical imputation, the DAE was best for divisions of 80-20 and 60-40 with a 70% missingness. While for 40-60 and 20-80 data divisions, categorical imputation using MissForest was better compared to DAE and MICE imputation. For the adult dataset, the 80-20 and 60-40 data divisions showed that the DAE achieved better performance for both categorical and numerical imputations when the missingness was 70%. For 40-60 and 10% missingness, numerical imputation was still the best using DAE, but categorical imputation using MICE was better. For the 20-80 data division and 50% missingness, the best RMSE was achieved by the MissForest imputation.

In some cases, the baseline DAE imputation was better than the baseline MissForest; however, when GAN-based data was incorporated, the MissForest was better. For instance, the 60-40 division for the Australian dataset had a baseline RMSE for the MissForest of 993.38m, which was 790.32 for DAE at a missingness ratio of 70%. This shows that DAE imputation is better than MissForest. However, when our method was applied, the RMSEs were reduced to 677.16 and 788.42 for the MissForest and DAE, respectively. Similarly, the heart 80-20 division with a 10% missingness was best for the MissForest baseline for numerical and DAE for categorical imputation. However, after applying the proposed method, the lowest numerical RMSE was 5.645 using the MICE imputation compared to 6.000 for the MissForest baseline. Similarly, the MICE achieved a lower categorical RMSE of 0.1383 using TGAN compared to 0.1554 for the DAE baseline.

To compare the imputation results using each GAN, the German credit dataset showed that the performance of categorical imputation was better when TAGN was used in addition to the original training data with the baseline imputation algorithms. For numerical imputation, the TGAN resulted in the best RMSE when the missingness ratio was 70%; thus, the TAGN was overall better with the German credit dataset compared to CTGAN. Comparisons of the two GANs for the Australian credit dataset also show that the TGAN performed better than CTGAN. For the 80-20 data division, the CTGAN only performed well on the 10% missingness. The 60-40 CTGAN was better for the numerical imputation for 30 and 50% missingness ratios. The CTGAN was better for the 70% missingness in both categorical and numerical imputations. For other data divisions, while the baseline was the best, the TGAN performance was better than CTGAN. For the heart dataset, imputation using TGAN gave better missing value predictions compared to CTGAN for all the data divisions.
and missingness ratios. For the adult dataset, CTGAN was better than TGAN.

We compared the effect of oversampling on data imputation with each baseline method. Figure 3 shows that for all experiments using the MICE imputation with a total of 64 experiments. Imputation using TGAN for numerical data (TGAN_num) was good in 42 experiments, followed by the MICE baseline (MICE_num) in 15 experiments and CTGAN (CTGAN_num) in 11 experiments. Similarly, for categorical imputation using the MICE algorithm, the CTGAN-based oversampling (CTGAN_cat) performed well in 30 experiments followed by TGAN (TGAN_cat) in 16 experiments. As a comparison, the MissForest-based imputation for baseline (MissF_num and MissF_cat) and GAN-based imputation are represented in Figure 4, which shows that TGAN for categorical imputation using MissForest was best in 32 experiments followed by the MissForest baseline in 28 experiments. The performance of the CTGAN with MissForest was not improved compared to the baseline, which was comparable with the proposed method; however, the MissForest baseline did not perform well when the missingness was 70% in all four datasets. Finally, the comparison of the DAE baseline and GAN-based oversampling imputation is represented in Figure 5, which shows that when oversampling using TGAN followed by DAE to reconstruct missing values, TGAN (TGAN_cat) was best in 49 experiments for categorical imputation while the baseline method was only better in 12 experiments. Similarly, for numerical imputation, the TGAN was best in 37 experiments while the baseline method was best in only 14 experiments. This highly suggests that using the proposed method will improve the imputation performance of datasets with missing values.

Figure 6 represents the best results for the entire experiment, such as the best result obtained in each table. For example, Table 3 shows the best results obtained by TGAN for both categorical (TGAN_cat) and numerical (TGAN_num) imputation. Figure 4 shows the best results achieved in each experiment, where TGAN was best in most experiments in both categorical (best in 38 experiments) and numerical (best results in 31 experiments) imputation. Among the baseline methods, MissForest imputation was best in 5 and 19 experiments for numerical and categorical imputations, respectively, MICE for 6 and 0, and DAE for 12 and 1. Thus, MissForest was better than the other baseline methods. One reason for this is that all experiments were performed using default hyperparameters; however, MissForest does not require any parameter tuning, as described in [66]. Although we did not record the running times for each experiment explicitly, we found that MissForest is computationally complex compared to MICE and DAE. The computational complexity of MissForest is also evident from several studies [mention]. We also found that CTGAN performed better.
when the training dataset is larger. For smaller datasets, its performance was worse than TGAN. For the adult dataset, CTGAN was best in 7 experiments for categorical imputation and 5 for numerical imputation. The TGAN was the best in 4 experiments (for 20-80 divisions only) with numerical imputation and 6 experiments for categorical imputation. The DAE was better compared to other baseline methods on the adult dataset as it achieved better reconstruction errors in 7 experiments for numerical imputation, while the MissForest baseline method was best in 6 experiments for categorical imputation. Thus, the overall performance of CTGAN was best on the adult dataset. The results obtained from the detailed experiments highly suggest using the proposed method for better imputation of missing values. We used the recently proposed GANs and three different imputation models; however, algorithm 1 can easily be generalized to incorporate more GANs for data synthesis as well as baseline imputation models.

VI. CONCLUSION

Many algorithms for numerical data imputation have been proposed; however, algorithms to impute mixed datasets are scarce. In addition, currently available datasets contain a limited amount of data, while model-based imputation algorithms require sufficient complete training datasets. Therefore, this study proposed a method using GANs (TGAN and CTGAN) to synthesize samples from training data for better model training. We then used three different imputation models, including MICE, MissForest, and DAE, to impute missing values. We found from several experiments with different missingness ratios that our proposed method improves the performance of predicting missing values compared to the baseline methods. Although this study provides promising improvements compared to baseline methods, there are some limitations that we aim to address in future works.

We introduced random missingness into the data using the MCAR mechanism [14], [18], [67], [51]; however, future works will consider more experiments using different missingness criteria, such as MAR and MNAR [18]. This study only used mixed datasets because they are difficult to handle and represent both numerical and categorical data. However, to prove the effectiveness of our proposed model, future works will use purely categorical and numerical datasets. We also aim to generalize algorithm 1 for multiple GANs and baseline imputation models. Each model used in this study was implemented using default parameters, and no hyperparameter optimization was performed, which will be addressed in the future. We also aim to use the category information [51] to further improve the imputation performance.

APPENDIX

The supplementary materials are provided with the paper.

REFERENCES

[1] P. Liu, E. El-Darzi, L. Lei, C. Vasilakis, P. Chountas, and W. Huang, “An analysis of missing data treatment methods and their application to health care dataset,” in Proc. Int. Conf. Adv. Data Mining Appl., in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 3584, 2005, pp. 583–590, doi: 10.1007/11527503_69.

[2] M. A. Proschans, R. P. McMahon, J. H. Shih, S. A. Hansenber, N. L. Geller, G. Knatterud, and J. Wittes, “Sensitivity analysis using an imputation method for missing binary data in clinical trials,” J. Stat. Plan. Inference, vol. 96, no. 1, pp. 155–165, Jun. 2001, doi: 10.1016/S0378-3758(00)00353-3.

[3] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, “Pattern classification with missing data: A review,” Neural Comput. Appl., vol. 19, no. 2, pp. 263–282, Sep. 2010, doi: 10.1007/s00521-009-0295-6.

[4] L. Qu, J. Hu, L. Li, and Y. Zhang, “PPCA-based missing data imputation for traffic flow volume: A systematic approach,” IEEE Trans. Intell. Transp. Syst., vol. 10, no. 3, pp. 512–522, Sep. 2009, doi: 10.1109/TITS.2009.2026312.

[5] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE J. Sel. Topics Signal Process., vol. 4, no. 2, pp. 272–287, Apr. 2010, doi: 10.1109/JSTSP.2009.2039171.

[6] C. Ji and A. Elwalid, “Measurement-based network monitoring: Missing data formulation and scalability analysis,” in Proc. IEEE Int. Symp. Inf. Theory, Jun. 2000, p. 78, doi: 10.1109/ISIT.2000.866368.

[7] M. Osman, A. M. Abu-Mahfouz, and P. R. Page, “A survey on data imputation techniques: Water distribution system as a use case,” IEEE Access, vol. 6, pp. 63279–63291, 2018, doi: 10.1109/ACCESS.2018.2877269.

[8] J. W. Graham, “Missing data analysis: Making it work in the real world,” Annu. Rev. Psychol., vol. 60, no. 1, pp. 549–576, Jan. 2009, doi: 10.1146/ANNUREV.PSYCH.58.110405.085530.

[9] 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,” Artif. Intell. Med., vol. 50, no. 2, pp. 105–115, Oct. 2010, doi: 10.1016/j.artmed.2010.05.002.

[10] K. Lakshiminarayan, S. A. Harp, R. Goldman, and T. Samad. (1996). Imputation of Missing Data Using Machine Learning Techniques. [Online]. Available: https://www.aaai.org

[11] D. Bertsimas, C. Pawlowski, and Y. D. Zhao, “From predictive methods to missing data imputation: An optimization approach,” J. Mach. Learn. Res., vol. 18, no. 196, pp. 1–39, 2018.

[12] L. Gondara and K. Wang, “MIDA: Multiple imputation using denoising autoencoders,” in Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10939, Jun. 2018, pp. 260–272, doi: 10.1007/978-3-319-93040-4_21.
C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation,” Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 1096–1103, doi: 10.1145/1390156.1390294.

S. J. Choudhury and N. R. Pal, “Imputation of missing data with neural networks for classification,” Knowl.-Based Syst., vol. 182, Oct. 2019, Art. no. 104838, doi: 10.1016/j.knosys.2019.07.009.

R. Lall and T. Robinson, “The MIDAS touch: Accurate and scalable missing-data imputation with deep learning,” Political Anal., vol. 30, no. 2, pp. 179–196, 2022, doi: 10.1017/PAN.2020.49.

A. Ahmad and S. S. Khan, “InitKmix—A novel initial partition generation algorithm for clustering mixed data using k-means-based clustering,” Expert Syst. Appl., vol. 167, Apr. 2021, Art. no. 114149, doi: 10.1016/j.eswa.2020.114149.

A. Ahmad and S. S. Khan, “Survey of state-of-the-art mixed data clustering algorithms,” IEEE Access, vol. 7, paper 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

J. L. Schafer, Analysis of Incomplete Multivariate Data. Boca Raton, FL, USA: CRC Press, Aug. 1997.

S. S. Khan, A. Ahmad, and A. Mihailidis, “Bootstrapping and multiple imputation ensemble approaches for missing data,” Oct. 2019, arXiv:1802.00154. Accessed: Dec. 30, 2021.

J. N. K. Rao and Q. Wang, “Empirical likelihood-based inference under imputation for missing response data,” Ann. Statist., vol. 30, no. 3, pp. 896–924, Jun. 2002, doi: 10.1214/aos/1028674845.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, Jun. 2001, doi: 10.1093/bioinformatics/17.6.520.

T. D. Pigott, “A review of methods for missing data,” Educ. Res. Eval., vol. 7, no. 4, pp. 353–383, Dec. 2001, doi: 10.1177/1356238104266950.

S. Nijman, A. Leeuwengen, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, “Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review,” J. Clin. Epidemiol., vol. 142, pp. 218–229, Feb. 2022, doi: 10.1016/j.jclinepi.2021.11.023.

W.-C. Lin, C.-F. Tsai, and J. R. Zhong, “Deep learning for missing data imputation: A novel centroid initialization in missing value imputation towards mixed datasets,” Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), 2018, pp. 289–293, doi: 10.1109/ISBI.2018.8363576.

L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modelling tabular data using conditional GAN,” Oct. 2019, arXiv:1907.00503. Accessed: Dec. 29, 2021.

T. Siwawatting, T. Anwar, D. Sarwinda, and H. S. Al-Ash, “A novel centroid initialization in missing value imputation towards mixed datasets,” Commun. Math. Biol. Neurosci., vol. 2021, no. 2021, Dec. 2021, Art. no. 11, doi: 10.28919/cmbn/5344.

C.-F. Tsai, M.-L. Li, and W.-C. Lin, “A class center based approach for missing value imputation,” Knowl.-Based Syst., vol. 151, pp. 124–135, Jul. 2018, doi: 10.1016/j.knosys.2018.03.026.

Z. Zhang, “Missing data imputation: Focusing on single imputation,” Ann. Transl. Med., vol. 4, no. 1, p. 9, Jan. 2016, doi: 10.29733/j.issn.2056-5839.2015.12.38.

X. Zhang, X. Song, H. Wang, and H. Zhang, “Sequential local least squares imputation estimating missing value of microarray data,” Comput. Biol. Med., vol. 38, no. 10, pp. 1112–1120, Oct. 2008, doi: 10.1016/j.compbiomed.2008.08.006.

A. R. Alsaheer, J. Pan, and A. Al-Hurban, “Handling complex missing data using random forest approach for an air quality monitoring dataset: A case study of Kuwait environmental data (2012 to 2018),” Int. J. Environ. Res. Public Health, vol. 18, no. 3, pp. 1–26, Feb. 2021, doi: 10.3390/IJERPH18031333.

A. Aleyani, W. Wang, and B. de la Iglesia, “Multiple imputation ensembles (MIE) for dealing with missing data,” Social Netw. Comput. Sci., vol. 1, no. 3, p. 134, Apr. 2020, doi: 10.1016/j.socnet.2020.05.001.

S. Awawdeh, H. Fariis, and H. Hiary, “Evompute: An evolutionary approach for missing data imputation and feature selection in the context of supervised learning,” Knowl.-Based Syst., vol. 236, Jan. 2022, Art. no. 107734, doi: 10.1016/j.knosys.2021.107734.

UCI Machine Learning Repository: Adult Data Set. Accessed: Dec. 30, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult

Cooperative Election Study. Accessed: Dec. 30, 2021. [Online]. Available: https://cces.gov.harvard.edu/home

Y. Wang, D. Li, X. Li, and M. Yang, “PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data,” Apr. 2021, arXiv:2101.07770. Accessed: Dec. 30, 2021.

M. Pampuca, G. Hutcheson, and J. Williams, “Handling missing data: Analysis of a challenging data set using multiple imputation,” Int. J. Res. Method Educ., vol. 39, no. 1, pp. 19–37, Jan. 2016, doi: 10.1080/1743727X.2014.979146.

Y. Zhao and M. Udell, “Missing value imputation for mixed data via Gaussian copula,” in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, vol. 11, Aug. 2020, pp. 636–646, doi: 10.1145/3394486.3403106.

G. B. Durrant, “Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates,” Int. J. Social Res. Methodol., vol. 12, no. 4, pp. 293–304, Oct. 2009, doi: 10.1086/53439003.

S. Jäger, A. Allhorn, and F. Bießmann, “GAN synthesis and composing robust features with denoising autoencoders,” in Proc. 25th Int. Conf. Mach. Learn. (ICML), 2008, pp. 497–504, doi: 10.1109/ICML.2008.203.

A. Purwar and S. K. Singh, “Hybrid prediction model with missing value imputation,” Expert Syst. Appl., vol. 42, no. 13, pp. 5621–5631, Aug. 2015, doi: 10.1016/j.eswa.2015.02.050.

C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, “GAN augmentation: Augmenting training data using generative adversarial networks,” 2018, arXiv:1810.10863.
and the Chancellor’s Innovation Award in technology, in 2015. awards such as the University Merit Awards, in 2020 and 2022, the College nition for a valuable Ph.D. work, in 2004. He received several scholarship biology, healthcare, and education. He was a recipient of the Dean’s Recog- on developing intelligent algorithms to solve problems in domains such as artificial intelligence, data mining, and machine learning. He mainly focuses analytics research and services. He has published more than 120 scientific machine learning algorithms for various medical and civil applications.

T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” Mar. 2018, arXiv:1812.04948. Accessed: Dec. 30, 2021.

Z. Lin, A. Khetan, G. Fanti, and S. Oh, “PacGAN: The power of two samples in generative adversarial networks,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 1–10.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of Wasserstein GANs,” Dec. 2017, arXiv:1704.00028. Accessed: Dec. 29, 2021.

M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: What is it and how does it work?” Int. J. Methods Psychiatr. Res., vol. 20, no. 1, pp. 40–49, Mar. 2011, doi: 10.1002/MPR.329.

D. J. Stekhoven and P. Bühlmann, “MissForest—Non-parametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/BIOINFORMATICS/BTR597.

J. Poulos and R. Valle, “Missing data imputation for supervised learning,” Appl. Artif. Intell., vol. 32, no. 2, pp. 186–196, Apr. 2018, doi: 10.1080/08839514.2018.1448143.

WASIF KHAN received the master’s degree from Kasetsart University, Bangkok, Thailand, under TAIST Tokyo Tech Program in collaboration with the Tokyo Institute of Technology Japan, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University (UAEU). His research interests include computer vision, deep learning, and developing explainable intelligent machine learning algorithms for various medical and civil applications.

MOHAMMAD MEHEDY MASUD (Member, IEEE) received the Ph.D. degree in computer science from the University of Texas at Dallas, USA, in December 2009. He is currently an Associate Professor at the College of Information Technology (CIT), United Arab Emirates University (UAEU). He has published more than 60 research articles in impact factor journals including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Knowledge and Information Systems (KAIS) journal, and high ranked conferences including the IEEE International Conference on Data Mining (ICDM). His is the lead author of the book Data Mining Tools for Malware Detection and the principal inventor of a U.S. Patent. His research interests include machine learning, big data mining, healthcare data analytics and decision support, data stream mining, and intrusion detection using data mining.

LUQMAN ALI received the B.Sc. degree in electrical communication engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2014, and the master's degree from Kasetsart University, Thailand, under TAIST TOKYO TECH Program in collaboration with the Tokyo Institute of Technology, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University. His research interests include computer vision and deep learning. He mainly focuses on developing intelligent machine learning algorithms for various medical and civil applications. His awards and honors include the TAIST Tokyo Scholarship (Kasetsart University), Soyama Scholarship (Tokyo Institute of Technology), Sakura Exchange (Tokyo institute of Technology and Full Fellowship from United Arab Emirates University).

NASLOON ALI received the Ph.D. degree from the College of Medicine and Health Sciences, Institute of Public Health. She is currently a Post-doctoral Research Fellow with the College of Medicine and Health Sciences, Institute of Pub- lic Health. She is also an Epidemiologist with interests in maternal and child health as well as social and behavioral public health. Her current research interests include pregnancy loss, infertility, adverse maternal outcomes, and mental health during pregnancy.

LUAI A. AHMED is currently an Epidemiolo- gist with special interests in the epidemiology and prevention of noncommunicable chronic diseases. He has developed, performed and coordinated suc- cessful research projects and published several scientific articles. His current research interests include epidemiology of diabetes, cancer, and the mother and child health. He is the PI for the “Mutaba’al” Mother and Child Health Study.

LUAI A. AHMED received the Ph.D. degree in computer science from The University of Manchester, U.K. He is currently an Associate Professor with the College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates. His research interests include machine learning, data mining, and nanotechnology.

AMIR AHMAD received the Ph.D. degree in computer science from the University of Texas at Dallas, USA, in December 2009. He is currently an Associate Professor at the College of Information Technology (CIT), United Arab Emirates University (UAEU). He has published more than 60 research articles in impact factor journals including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Knowledge and Information Systems (KAIS) journal, and high ranked conferences including the IEEE International Conference on Data Mining (ICDM). His is the lead author of the book Data Mining Tools for Malware Detection and the principal inventor of a U.S. Patent. His research interests include machine learning, big data mining, healthcare data analytics and decision support, data stream mining, and intrusion detection using data mining.

VOLUME 10, 2022

[58] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” Jan. 2016, arXiv:1511.06434. Accessed: Dec. 30, 2021.

[59] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” Mar. 2018, arXiv:1812.04948. Accessed: Dec. 30, 2021.

[60] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” Nov. 2014, arXiv:1411.1784. Accessed: Dec. 30, 2021.

[61] J. Engelmann and S. Lessmann, “Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning,” Expert Syst. Appl., vol. 174, Jul. 2021, Art. no. 114582, doi: 10.1016/ESWA.2021.114582.

[62] L. Xu, “Synthesizing tabular data using conditional GAN,” M.S. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2020. Accessed: Dec. 29, 2021. [Online]. Available: https://dspace.mit.edu/handle/1721.1/128349

[63] Z. Lin, A. Khetan, G. Fanti, and S. Oh, “PacGAN: The power of two samples in generative adversarial networks,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 1–10.

[64] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of Wasserstein GANs,” Dec. 2017, arXiv:1704.00028. Accessed: Dec. 29, 2021.

[65] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: What is it and how does it work?” Int. J. Methods Psychiatr. Res., vol. 20, no. 1, pp. 40–49, Mar. 2011, doi: 10.1002/MPR.329.

[66] D. J. Stekhoven and P. Bühlmann, “MissForest—Non-parametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/BIOINFORMATICS/BTR597.

[67] J. Poulos and R. Valle, “Missing data imputation for supervised learning,” Appl. Artif. Intell., vol. 32, no. 2, pp. 186–196, Apr. 2018, doi: 10.1080/08839514.2018.1448143.

Nazar Zaki graduated from the Universiti Teknologi Malaysia (UTM). He served as the Chair of the Department of Computer Science and Software Engineering, CIT, United Arab Emirates University (UAEU), for almost ten years introduced new academic programs and contributed significantly to the establishment and success of the department. He is currently a Professor of computer science. He is the Founder and the formal Director of the Big Data Analytics Center, UAEU, which aims to ingrain a sustained impact through groundbreaking data analytics research and services. He has published more than 120 scientific results in reputable journals and conferences. His research interests include artificial intelligence, data mining, and machine learning. He mainly focuses on developing intelligent algorithms to solve problems in domains such as biology, healthcare, and education. He was a recipient of the Dean’s Recognition for a valuable Ph.D. work, in 2004. He received several scholarship awards such as the University Merit Awards, in 2020 and 2022, the College Recognition Award for Excellence in Scholarship, in 2007, 2012, and 2016, and the Chancellor’s Innovation Award in technology, in 2015.