Rigorous Experimental Analysis of Tabular Data Generated using TVAE and CTGAN

Parul Yadav, Manish Gaur, Rahul Kumar Madhukar, Gaurav Verma, Pankaj Kumar,
Nishat Fatima, Saqib Sarwar, Yash Raj Dwivedi
Computer Science and Engineering Department,
Institute of Engineering and Technology,
Lucknow, UP 226021, India

***Abstract*—Synthetic data generation research has been progressing at a rapid pace and novel methods are being designed every now and then. Earlier, statistical methods were used to learn the distributions of real data and then sample synthetic data from those distributions. Recent advances in generative models have led to more efficient modeling of complex high-dimensional datasets. Also, privacy concerns have led to the development of robust models with lesser risk of privacy breaches. Firstly, the paper presents a comprehensive survey of existing techniques for tabular data generation and evaluation matrices. Secondly, it elaborates on a comparative analysis of state-of- the-art synthetic data generation techniques, specifically CTGAN and TVAE for small, medium, and large-scale datasets with varying data distributions. It further evaluates the synthetic data using quantitative and qualitative metrics/techniques. Finally, this paper presents the outcomes and also highlights the issues and shortcomings which are still need to be addressed.**

***Keywords*—Synthetic data generation; tabular data generation; data privacy; conditional generative adversarial networks; variational autoencoder**

I. Introduction

Tabular Data made up of databases with tabular structures consisting of rows representing observations and columns representing features. In the digital era, "data" is considered as "new water". On the one side,compliance of privacy laws like GDPA has imposed an obligation on organizations to secure and protect private and sensitive data,while on the other side, data acts as a fuel in the wide range of machine learning applications like Cloud migration, Artificial Intelligence (AI)/ ML model training, application testing, simulation analysis, data sharing, scientific trials, and new product development. An innovation solution to address the issue is to generate the synthetic data.

Synthetic data is an artificially generated data which carries the same statistical properties (i.e. mean, median, mode, correlation, regression, etc.). as the real data. Synthetic data is useful where it is challenging to obtain and use real data due to privacy concerns and difficulty in collecting real data, augmenting small datasets, and range of machine learning applications. Moreover, real data can be of types like tabular, time-series, audio, video, medical images/ signals etc. Out of these types, tabular data is the most commonly used form of data and generation of synthetic data for it is still challenging and requires to address the multiple constraints/ characteristics of tabular data to produce quality synthetic data.

There are multiple constraints (multimodal, class imbalance, non Gaussian data, learning from sparse one-hot-encoded vectors, mixed data type) inherently present in the tabular data that need to be addressed while generating its synthetic counterpart. Along with these constraints, a feature has varying characteristics. In this paper, tabular data features have been categorized into four categories: continuous, categorical, mixed type, and anonymized. Characteristics of each tabular data feature are shown in Table I. Generation of synthetic tabular data requires identification of the distributions of each feature and simultaneously mapping the correlations present in the data.

TABLE I. Data Characteristics

Feature Type	Details
Continuous Columns	Gaussian Distribution Multi-Modal Distribution Long Tail Distribution
Categorical Columns	Binary Categorical Multiple Categorical
Mixed Type Columns	Missing Values A high-frequency finite value mixed with other values
Anonymized Columns	Unique Columns - Roll Number, Patient ID, etc. Non-Unique Columns - Name, Country, etc.

Broadly, there are two categories of models used to generate synthetic tabular data, namely, statistical or Probability [4], [5] and machine learning [7], [8], [9], [10] methods based models.

Probabilistic models have existed and are continually being developed to generate synthetic data. But recent advancements in deep learning-based generative models, especially GANs [1], have made them state-of-the-art in the field of synthetic data generation. Moreover, the existence of stricter privacy laws and general awareness regarding user privacy have made data sharing difficult. This has led to the development of novel privacy-preserving mechanisms to ease the generation and sharing of data. Differential privacy [2] has become the gold standard in this domain. It uses randomized algorithms [3] for the sanitization of sensitive information and also limits the privacy risk of revealing sensitive information.

Besides the generation of synthetic data, robust evaluation mechanisms [2], [11] are of equal importance. This paper presents exciting methods for synthetic data generation and

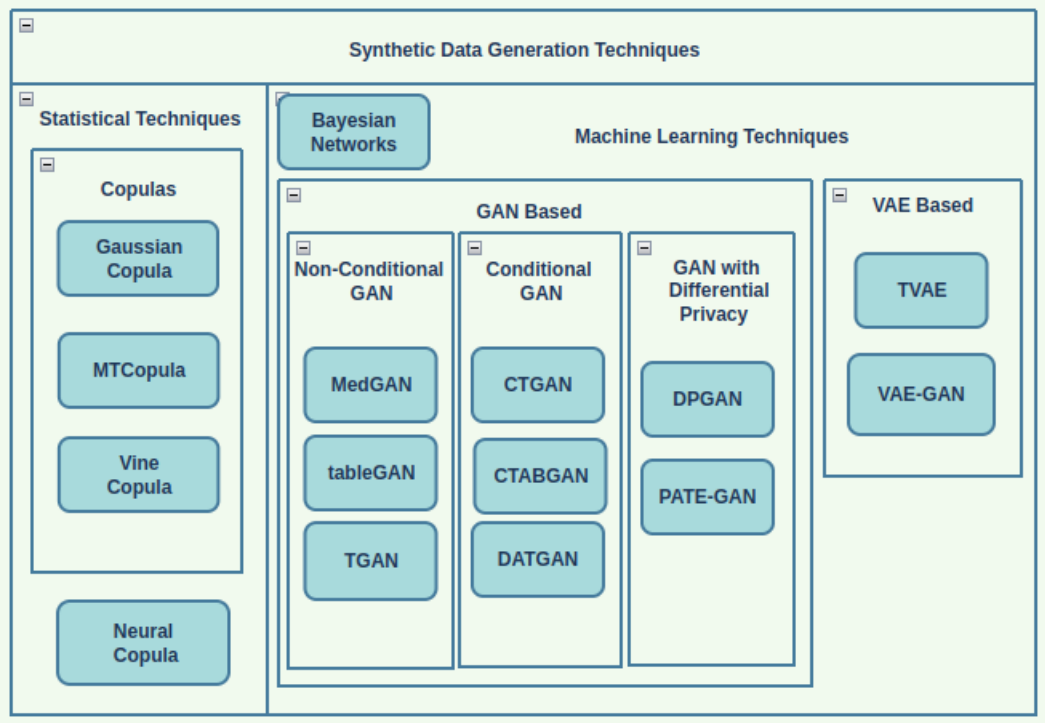

Fig. 1. Synthetic data generation techniques.

elaborates on evaluation mechanisms. This paper highlights rigorous analysis to analyze the performance of the state-of-the-art models for tabular data generation on varied nature data. Presented is a rigorous comparative analysis of the state-of-the-art models, namely, TVAE [12] and CTGAN [12] for small-*iris* [13] and *breast cancer* [13], medium-*adult* [13] and large-*credit* [13] datasets with comparative statistical scores and visualization reports.

The structure of the paper is organized as follows: Section II describes existing methods for synthetic data generation extensively; Section III lists the techniques for evaluation of generated synthetic data in detail; Section IV describes the datasets, models, and evaluation metrices used in the research; Section V provides an in-depth analysis of the results obtained; and Section VI presents a short qualitative summary of the proposed research in the conclusion.

II. Existing Models for Synthetic Tabular Data Generation

The existing models for generating synthetic tabular data have been broadly classified into two categories namely, statistical methods-based models [4], [5], [6] and machine learning-based techniques [7], [8], [9], [10] as shown in the Fig. 1. This section entails a detailed and comprehensive in-depth analysis of each methods at hand.

A. Statistical Methods-based Models

Several statistical methods-based models [4], [5], [6] have existed, and novel methods have been proposed to address the task of synthetic data generation. One of the earliest statistical techniques for synthetic data generation is Inverse Transform Sampling [6] which involves sampling data from a known data distribution for the random variable X. It generates independent univariate samples. Thereafter, a perturbation technique involving fitting a multivariate Gaussian distribution on input data is introduced. The General Additive Data Perturbation(GADP) technique [4] generated synthetic data by adding a noise variable to the estimated distribution. Another variant of GADP is the Dirichlet Multivariate Synthesizer [6] which is based on Maximum Likelihood Estimation (MLE) [14]. The problem with MLE is that the computation increases exponentially as the number of variables increases. Apart from these statistical methods, one of the most useful statistical methods for synthetic data generation uses copulas. Details of copulas are described in Sub-section II-A1

1) Copulas: A copula [15] is a mathematical function that describes the correlation between the marginals of random variables. This helps in identifying the multivariate joint distribution for a set of random variables. A lot of research has been done to identify the right parameters for the copula model and the marginals. Based on these, several variants of copula have been proposed. Gaussian Copula [5] is the most popular and one of the very few copula functions available for modeling the joint multivariate probability distribution. Apart from the Gaussian Copula [5] and t-Copula [16] models for multivariate distribution, Vine Copula [17] models have gained prominence lately as a modeling method as they are built only on the univariate and bivariate distributions. More recently, neural network techniques are being incorporated to identify the right set of parameters to construct a generic copula that models any multivariate joint distribution [7].

a) MTCopula: Since Gaussian Copula [5] fails to address the complex distributions of marginals and the joint

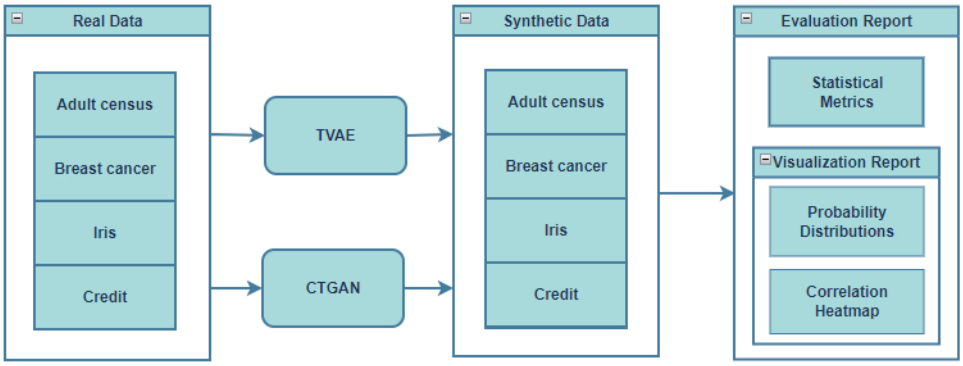

Fig. 2. Proposed comparison model.

distribution, Benali et al. propose a novel MTCopula [6] which involves parameter selection criteria for selecting the parametric marginal and multivariate copula functions based on Akaike Information Criterion (AIC) [18]. The main drawback of this approach is that it selects one of the existing parametric marginals or parametric copulas. The complex distributions of a random variable may not be exactly modeled by parametric marginals. Moreover, it considers only Gaussian Copula [5] and t-Copula [16] for the selection of multivariate copula functions, and the complex relationships among the random variables are not always captured by these two generic copula.

b) Vine copula: Vine copulas [17] are a special class of copula models as they use bivariate distribution as the building blocks for multivariate joint distribution [15]. This is achieved by forming a vine-like structure, one node at a time. However, with the increase in the number of variables, the number of feasible configurations of a vine copula expands exponentially, making model selection a significant development issue. This problem is addressed by [9] using reinforcement learning and selecting tree levels sequentially while using LSTM networks [19] to learn from vine configurations.

B. Machine Learning Methods-based Models

Recent advancements in the development of generative models [1] have transformed visual media-centric research to new heights. The ability of generative models is now being utilized to learn the complex relationships of tabular data and generate similar synthetic data. The two most important techniques using generative models are variational autoencoders [20] and generative adversarial networks [1], which are discussed in the following section.

1) Bayesian Networks: Bayesian networks [21] are probabilistic graphical models used to determine probabilistic inferences between variables. They are frequently used in computational systems biological method [22] to understand the underlying biological relationships. Here, the dependencies of variables are defined prior to training. This is also a major drawback, as it requires prior information on the dataset. Moreover, it becomes computationally expensive when dealing with large and sparse datasets.

2) Variational Autoencoders: The architecture of VAEs [20] consists of an encoder network and a decoder network. The encoder takes real data as input and converts it into a vector corresponding to a latent distribution. This vector is served as an input to the decoder, which reconstructs the real data that was served as an input to the encoder. On training completion, the decoder network can now generate new samples. The variational part brings randomness to this process.

a) Differentially private autoencoder: [23] introduced a novel differentially private autoencoder for synthetic data generation.

b) TVAE: Xu et al [12] propose a novel VAE known as TVAE for tabular data using two neural networks, one for the encoder and the other for the decoder network, and train them using Evidence Lower-Bound (ELBO) loss [24].

3) Generative Adversarial Networks: GAN [1] is based on the adversarial training of the generator network and the discriminator network, where the task of the generator is to produce fake samples that closely resemble the real samples while the discriminator tries to distinguish between the real and fake samples. GANs also belong to a generative class of models, but the key distinction between GANs and VAEs is that in VAEs, the encoder sees the real data, while in GAN, real data is not visible to the generator network. This is particularly useful in privacy-oriented applications. Different types of GANs are discussed below.

a) Non-Conditional GAN: GANs were first incorporated for synthetic tabular data generation in MedGAN [25], Table-GAN [26] and TGAN [8]. Vanilla GAN [1] architectures usually suffer from the problem of vanishing gradients, which leads to mode collapse. This has led to the rise of several variations of GAN architectures, such as WGAN, Wasserstein GAN [27] and conditional GANs [28].

b) Conditional GAN: When allowing the GAN [1] model to condition external information, it can generate samples by operating in different modes based on the contextual information provided. Thus, conditional GAN [28] is an extension of the GAN [1] architecture with the conditional operation. The different variations of conditional GANs are as follows.

- *CTGAN* [12] deals with problems like mixed data types, multimodal distributions, and imbalanced categorical columns of tabular data extensively. For the

problem of multimodal distribution in continuous data, it fits a variational Gaussian Mixture Model [29]. Thereafter, a mode is sampled from the identifiable modes, and the column is normalized based on the probability density of the selected mode. For discrete columns, one hot encoding is used.
The problem of adequately representing all the categories in the synthetically generated data is handled using a conditional vector. Since the minority category may not be adequately represented in the synthetic data, instead of providing random noise to the generator, it uses a conditional vector constrained on each respective category to train the generator and discriminator network. Finally, the discriminator is trained using a sampling method based on the conditional vector, i.e., sampling a row from the real data constraining each respective category.

- *CTABGAN* Conditional Tabular GAN (CTAB-GAN) [10] is based on the Convolutional Neural Network with an additional classifier network based on a multilayer perceptron apart from the generator and discriminator networks. It considers not only continuous and categorical variables but also a third class of variables known as mixed variables and also deals with variables with long-tail distributions. Zhao et al. proposed an advanced version of CTABGAN, CTABGAN+ [30] using Wasserstein loss [27] with gradient penalty and training with differential privacy stochastic gradient descent to ensure strict privacy guarantees.
- *DATGAN* [31] proposes DATGAN which is a novel architecture based on GAN using Directed Acyclic Graphs (DAGs) to model the information about the dataset. It uses LSTM cells to model expert knowledge using DAG.

c) GAN with Differential Privacy: Jordan et al. [32] apply Private Aggregation of Teacher Ensembles (PATE) [33] to the GAN model so as to obtain a generative model with tight differential privacy guarantees. Xie et al. [34] add noise to gradients to achieve differential privacy with GANs. Evaluation techniques that are utilized to assess the quality of generated synthetic data are explained below.

III. Evaluation Techniques

Synthetic data generation not only focuses on generation algorithms but also on robust evaluation mechanisms that can highlight the quality of generated synthetic data. To this end, evaluation mechanisms are classified into broadly three categories (quantitative, qualitative, and machine learning utility), highlighting each unique aspect of synthetic data. Further explanation is provided on these categories along with privacy preservability and differential privacy.

A. Quantitative Statistical Similarity Measures

Ideally, the generated synthetic data should have the same statistical properties as the real data. To ensure this, several statistical tests and metrics were compared to the synthetic data with real data.

1) Kolmogorov-Smirnov Test: To measure the similarity between the data distribution of continuous columns in real data and the generated synthetic data, the two-sample Inverted Kolmogorov-Smirnov test [35], commonly referred to as KSTest, is utilized. The p-value and D statistic obtained represent the similarity between the column distributions. For the whole dataset, a mean of the D statistic is obtained considering all the continuous columns, and then it is subtracted from 1 to obtain the final score. The D statistic can be computed using the following equation:

$$D_{m,n} = max|F(x) - G(x)|. \tag{1}$$

where $F(x)$ is the cumulative distribution function of the first sample with size m and $G(x)$ is the cumulative distribution function of the second sample with size n.

2) Chi-Square Test: Similarly, for discrete data, the Chi-Square Test [36] is used. After applying it to all the discrete columns in the data, an average of the score is obtained.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where χ^2 = Chi Squared statistic, O_i = Observed value, E_i = Expected value

3) Wasserstein Distance: Also known as Earth Mover distance [37], it intuitively defines how much quantity should be transported from x to y to transform a probability distribution from P_A to P_B.

$$W := W(F_A, F_B) = (\int_0^1 |(F_A^{-1}(u) - F_B^{-1}(u)|^2 \, du)^{\frac{1}{2}} \tag{3}$$

Where, F_A^{-1} and F_B^{-1} are the corresponding quantile functions, and F_A and F_B are the associated cumulative distribution functions (CDFs).

4) Kullback-Leibler Divergence: Finally, the third metric considered for quantifying the difference between the two probability distributions is the Kullback-Leibler divergence [11] or KL divergence, encompassing both the continuous and discrete variants. For discrete probability distributions P and Q:

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{4}$$

For probability distributions P and Q of continuous variable:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x) log \frac{P(x)}{Q(x)} dx \tag{5}$$

B. Qualitative Visualization Techniques

Apart from the above-mentioned quantifiable measures, data visualization methods are useful for analyzing the quality of synthetic data. Of these, probability distributions for univariate and bivariate analysis and correlation heatmaps for multivariate analysis provide deep visual insights into the quality of the generated synthetic data.

C. Machine Learning Utility

One important aspect of the task of synthetic data generation is to produce synthetic data that has approximately the same machine-learning utility as the real data. Without any utility, the generated synthetic data might not be of any value.

D. Privacy Preservability

This is one of the most important aspects in the task of synthetic data generation, considering the present scenario. Strict laws (e.g., the European General Data Protection Regulation [38]) and privacy concerns play a major role in sharing sensitive data. To address these, various mechanisms have been proposed to ensure data privacy. These can be classified into two major techniques: distance-based metrics (DCR (Distance to Closest Record) [30] and NNDR (Nearest Neighbour Distance Ratio) [39], and differential privacy mechanisms.

1) DCR: A synthetic record's Euclidean distance from its nearest real neighbour is measured by DCR [30]. Ideally, the chance of a privacy breach decreases as DCR increases.

2) NNDR: The NNDR [39] measures the ratio between the Euclidean distance for the closest and second-closest actual neighbours to any matching synthetic record, as opposed to just measuring the closest neighbour. This ratio falls between 0 and 1. Better privacy is indicated by higher values. Sensitive information from the nearest real data record may be revealed by low NNDR values between synthetic and real data.

E. Differential Privacy

By reducing the impact of each data point based on a predetermined privacy budget, Differential Privacy [2] defends against privacy assaults. Renyi Differential Privacy (RDP) is used by Zhao et al. [30]. Because it sets more stringent limits on the privacy budget. RDP offers tighter limitations for tracking the cumulative privacy loss through a series of mechanisms using the composition theorem, making it a strictly stronger privacy definition than DP.

Up to this point, elaboration has been provided on existing methods for synthetic data generation and evaluation metrics for analyzing the quality of the generated data. The next section delves into the methodology for designing and implementing the comparison model.

IV. Methodology for Comparison Model

Comparing and analyzing two state-of-the-art synthetic tabular data generation techniques, namely, TVAE [12] and CTGAN [12], Four datasets were explored of varying sizes, data features, and characteristics. For qualitative analysis, a selection of three statistical metrics is made, which are then implemented and analyzed across both models using varying batch sizes (20, 50, 100, and 300) and epochs (100, 200, 500, and 5000). Additionally, a visualization report is generated using probabilistic distribution and correlation heatmaps for variables in the datasets.

The process flow highlighting each component in the proposed comparison model is shown in Fig. 2. It lists four real datasets to be used, which train the two models, TVAE [12] and CTGAN [12]. The trained models then generate synthetic samples for each dataset individually. Finally, this synthetic data, along with real data, is used to prepare an evaluation report that highlights quantitative statistical metrics and a qualitative visualization report depicting feature distributions and correlations for synthetic as well as real data. Further subsections describe the datasets (Section IV-A), algorithms (Section IV-B), and evaluation metrics (Section IV-C) used to design and implement the proposed comparison model in detail.

TABLE II. Datasets Details

Category	Name	Total Features	Total Records	Feature Distribution
small	iris	6	150	(5 continuous, 1 discrete)
	breast cancer	33	569	(32 continuous, 1 discrete)
medium	adult	15	32561	(6 continuous, 9 discrete)
large	credit	31	284807	(30 continuous, 1 discrete)

TABLE III. Credit Sample

	Total Records	Class 0	Class 1	% Class 0	% Class 1
Real	284807	284315	492	99.8273	0.1727
Sample (15%)	42721	42647	74	99.8268	0.1732

A. Datasets

Datasets were categorized based on their size into three broad categories: small (*iris* [13] and *breast-cancer* [13]), medium (*adult*) [13] and large (*credit*). Four standard datasets were considered in different domains with varied sizes and a mix of different variable types. The extensive diversity of the datasets is reflected in Table II. Table II shows the total features and their distribution as continuous or discrete features and complete records for each dataset. Moreover, for the *large* dataset, *credit*, approximately 280,000 records are available. A 15% sample of the original dataset is taken, as illustrated in Table III. Due to the high imbalance in the *credit* dataset, with just 0.173% of samples belonging to Class 1, the same imbalance ratio is maintained in the sample.

B. Models and Algorithms

Two state-of-the-art machine learning algorithms are compared across four different datasets. TVAE [12] and CTGAN [12] models are employed for all datasets, with hyperparameter tuning conducted for each. Hyperparameters are optimized based on the size of the dataset, as detailed in Table IV, Table V, Table VI, and Table VII.

C. Evaluation

Three quantitative statistical evaluation metrics were used, Chi-Square (CS Test), Inverted KS D Statistic (KS Test), and KL Divergence (KL_c for continuous and KL_d for discrete) for

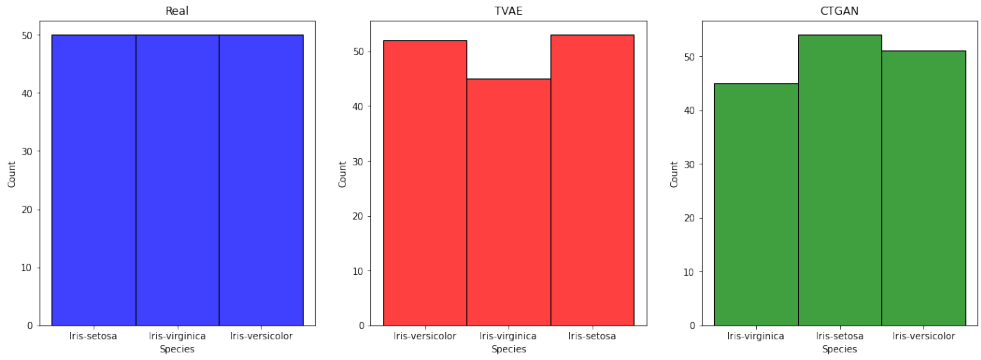

Fig. 3. Iris: Species Distribution.

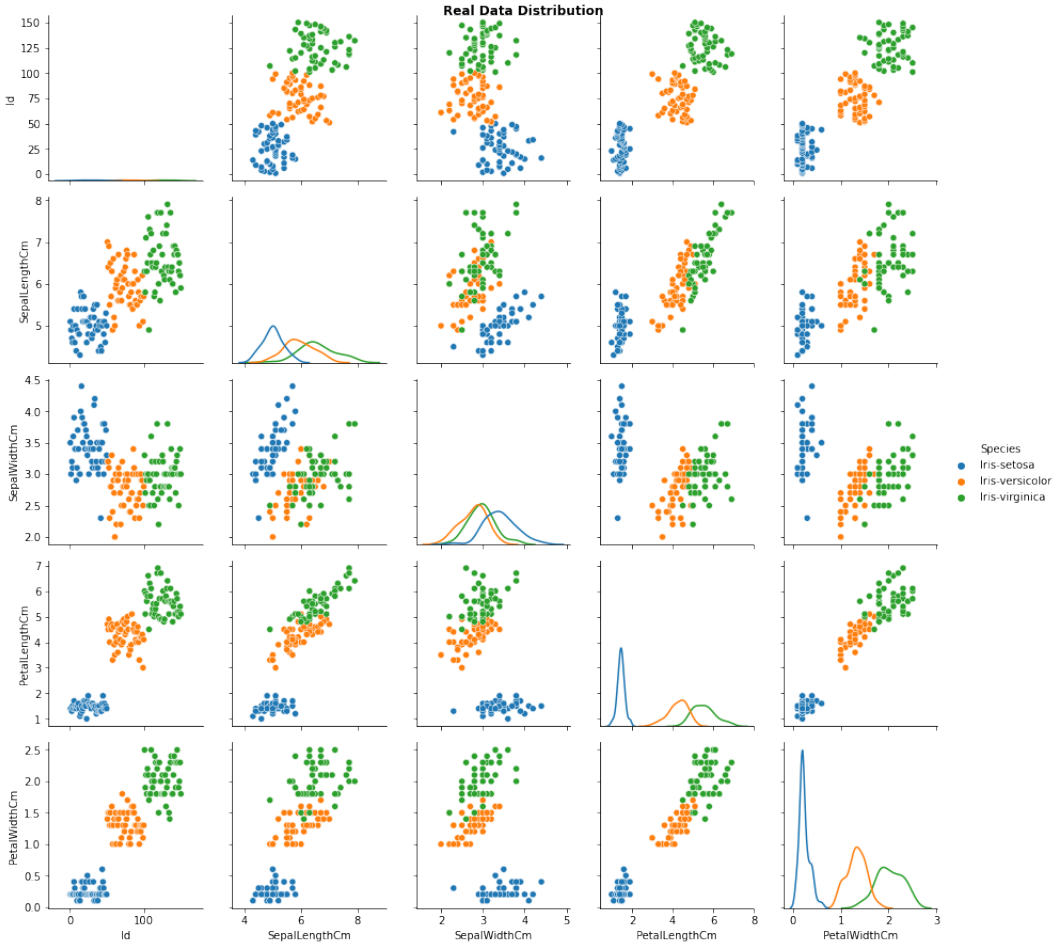

Fig. 4. Real Iris.

each category of the data, continuous or discrete (as applicable). Moreover, univariate distributions, bivariate distributions, and correlation heatmaps are utilized for qualitative data analysis through visualization. Categorical distributions for the *Iris* dataset are represented using histograms, as illustrated in Fig. 3. Further elaboration on these evaluation metrics can be found in Section III.

V. Results and Analysis

To analyze the results, the proposed comparison model was implemented in Python using its ML libraries. Details of the implementation environment are provided in Table VIII. TVAE [12] and CTGAN [12] models were implemented, trained for four datasets as explained in Subsection IV-A, and evaluated using three metrics outlined in Subsection IV-C, with variations in batch size and epochs, as shown in Table IV, Table V, Table VI, and Table VII. Statistical and visual observations

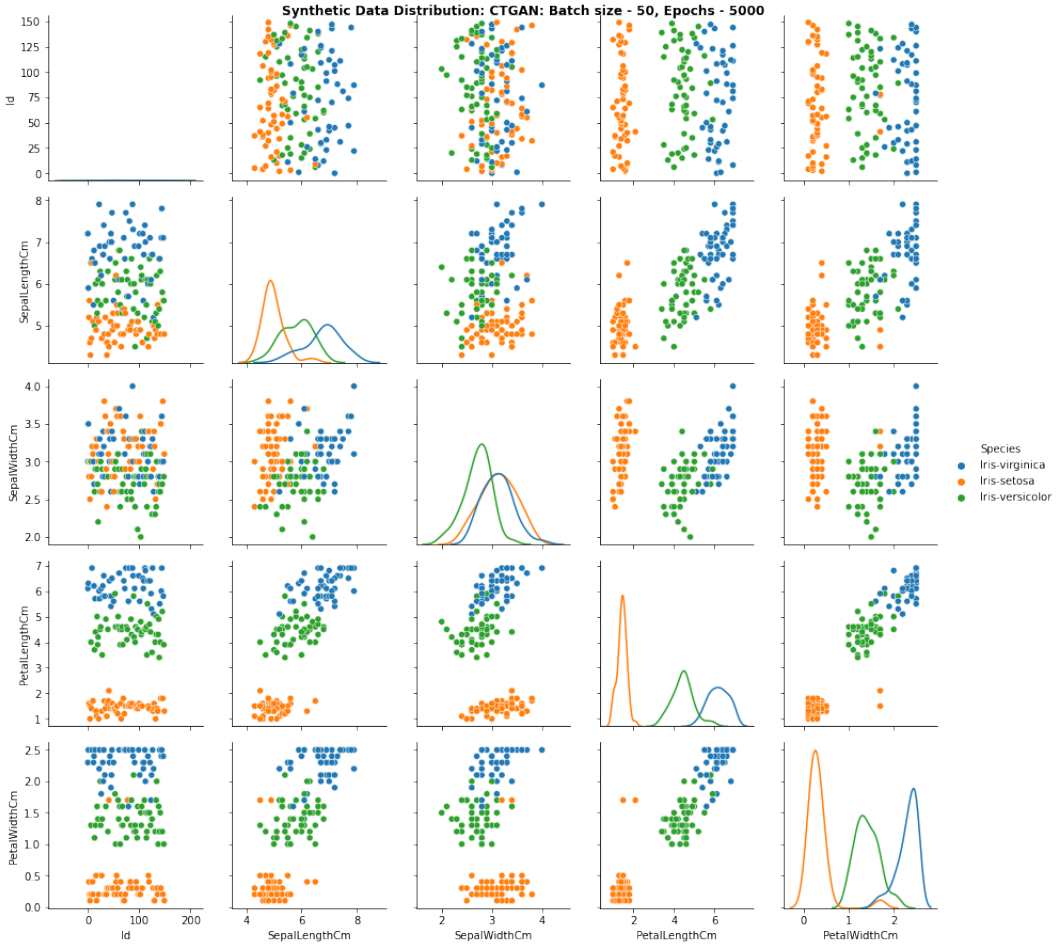

Fig. 5. CTGAN Iris.

were recorded, and the results were rigorously analyzed. This analysis will present the suitability as well as limitations of TVAE [12] and CTGAN [12] under varied and diversified features of the datasets, as shown in Table II.

In most of the cases, it appeared that the variational autoencoder-based TVAE [12] outperformed CTGAN [12], reflected by the higher KL divergence score, meaning a closer resemblance of the synthetic data distribution to the real data distribution. But the major advantage of the GAN [1] based model is the more effective privacy-preserving mechanism since the generator is unaware of real data values. These observations are rigorously analyzed in Subsection V-A and Subsection V-B.

A. Statistical Measures: Quantitative Evaluation

Based on the categories of the datasets, *small*, *medium*, and *large*, Batch sizes of the dataset were varied and described the statistical metrics obtained as follows.

For *small* datasets, *iris* and *breast-cancer*, the batch size is varied between 20, 50, and 100. Also, for *small* datasets, deep learning methods are more useful when training is done for longer periods of time, i.e., for larger epochs. Testing of this was done by training the data for 100, 500, and 5000 epochs, respectively.

From Table IV and Table V, observations were made that the KL divergence score is best for both *iris* and *breast-cancer* for the optimum value of 5000 epochs with a batch size of 50. This is true for the synthetic data generated using both TVAE [12] and CTGAN [12]. Another important observation from Table V is that change in the number of epochs and batch size did not have a significant influence on TVAE [12] while increasing the number of epochs for CTGAN [12] significantly increased all three metrics that were considered.

For the *medium* sized *adult* dataset, the batch size was increased to 300, and training was conducted for 200 and 300 epochs. Similar values were obtained for all metrics, as depicted in Table VI, with minimal significance observed in changing the epochs.

The credit dataset is sampled, preserving the minority-majority class ratio. The results obtained for a batch size of 300 and 200 epochs are shown in Table VII. The experiment is repeated three times, and the mean of all the values is shown in Table VII.

B. Visualization Analysis: Quantitative Evaluation

Apart from the statistical metrics, a visualization report with univariate and bivariate distributions and correlation

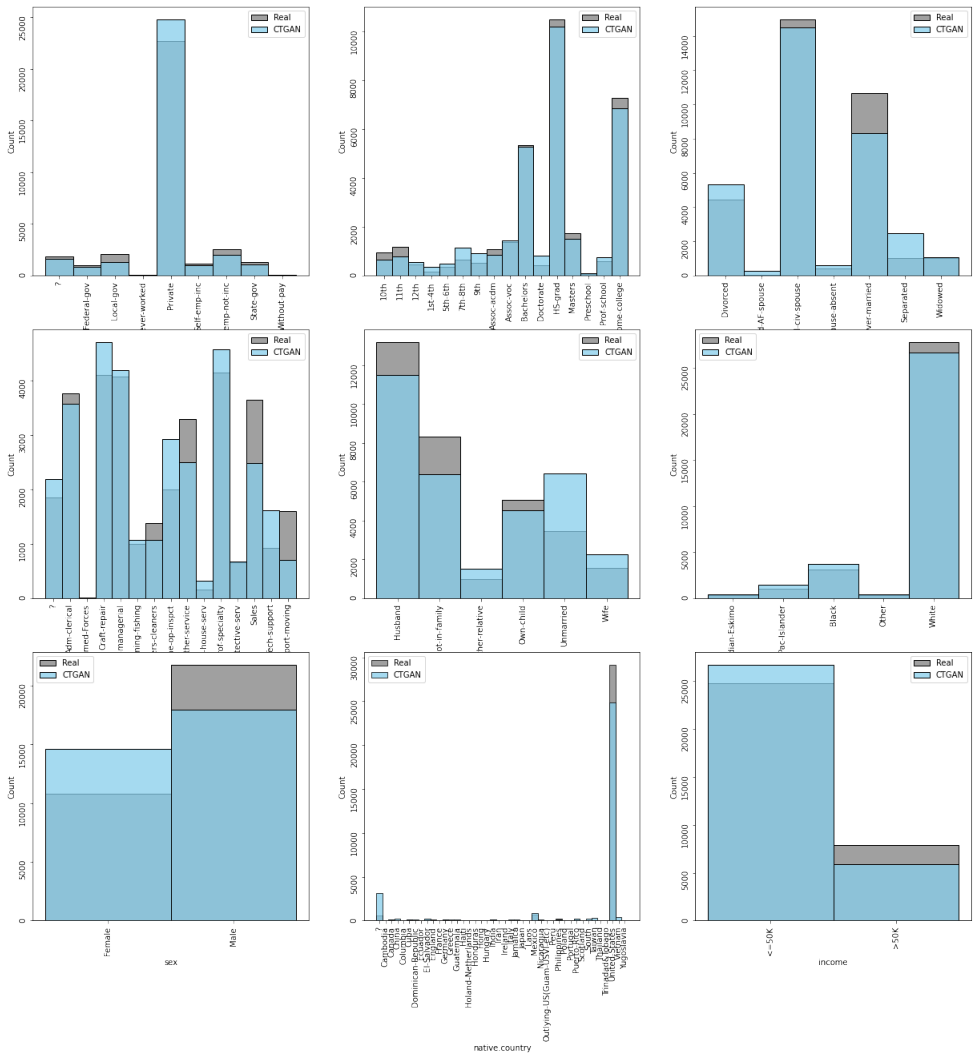

Fig. 6. Categorical distribution of real vs synthetic data.

heatmaps to study the similarity between real data and synthetic data for all four different sizes of datasets, namely, small (*iris* [13] and *breast-cancer* [13]), medium (*adult*) [13] and large (*credit*) was generated.

1) Probability Distributions: The univariate and bivariate distributions of variables based on the species type for the iris dataset were depicted in Fig. 4. Similar distributions are displayed for the synthetic version with the highest metric values for CTGAN [12] to analyze the statistical similarity, as shown in Fig. 5. Due to space constraints, only the distributions and heatmaps with the best statistical scores for synthetic data are presented. The distribution of the target variable "Species" is separately shown for real and synthetic data in Fig. 6. For the *breast-cancer* dataset, CTGAN [12] is good at capturing multi-modal distributions, as can be seen from Fig. 7. Univariate probability distributions for real and synthetic data (CTGAN) are shown in Fig. 7.

From Fig. 9 and Fig. 10, a comparative visualization of univariate and bivariate plots of numerical features for the real adult dataset and the synthetically generated dataset using CTGAN [12] is provided respectively. Additionally, the categorical histograms of the same real and synthetically generated data using CTGAN [12] are also depicted in Fig. 6. The categorical data histograms reflect the capability of the model to learn the minority and majority category distributions in a precise way.

Also, based on the univariate distributions obtained for the real and synthetic data as shown in Fig. 8, few observations that were made are:

The use of the variational Gaussian mixture model in CTGAN [12] oftentimes unnecessarily identified multiple modes as can be seen from the distribution plots of "V11", "V12", "V14", "V16" and "V17" in Fig. 8.

Modeling mix-type variables such as "Amount" and "V28" was a challenging task for both TVAE [12] and CTGAN [12] as the distribution got centered around zero, while a large number of finite values were not adequately represented.

Both TVAE [12] and CTGAN [12] failed to effectively capture and model the problem of class imbalance. As shown in Fig. 11, while TVAE [12] was unable to capture the minority

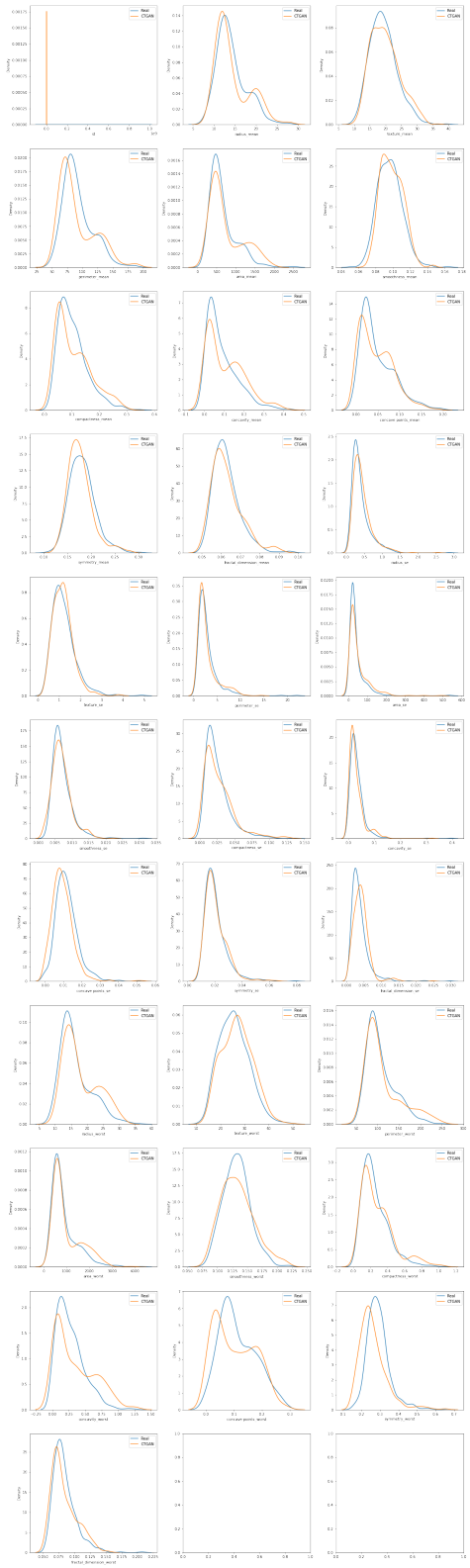

Fig. 7. Data Distribution: Breast Cancer - Real vs Synthetic (CTGAN).

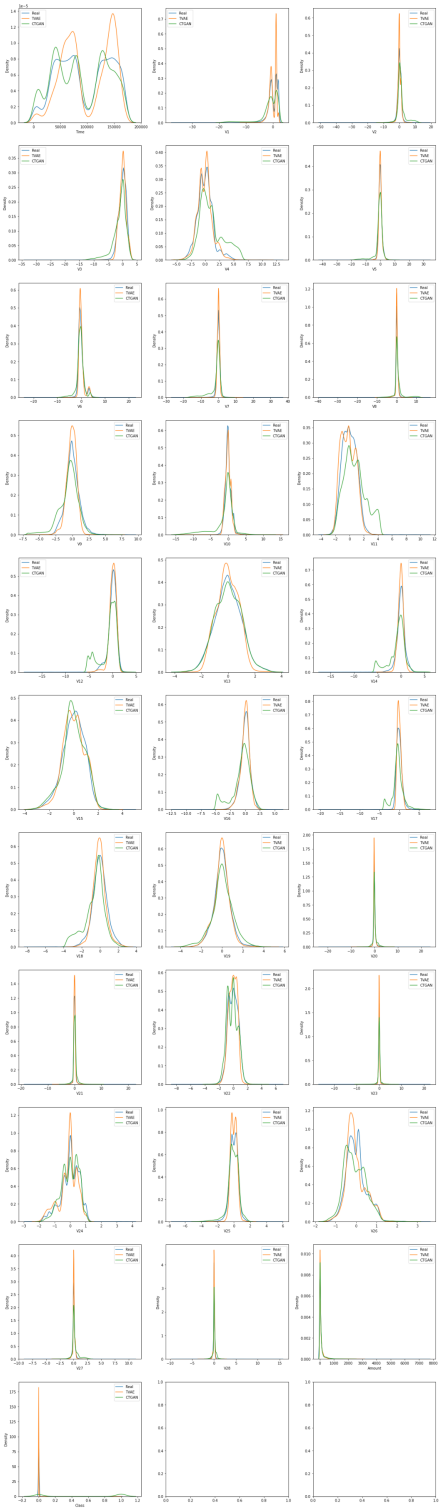

Fig. 8. Data Distribution: Credit - Real vs Synthetic.

class, CTGAN [12] oversampled the minority class in synthetic data, overriding the imbalance ratio.

Lastly, tuning the hyperparameters accordingly for each dataset improved the quality of the generated synthetic data for both TVAE [12] and CTGAN [12]. Thus, even high-dimensional data with complex relationships can be effectively modeled using generative models for synthetic data generation.

Fig. 9. Real Adult.

Fig. 10. CTGAN Adult.

2) Correlation Heatmaps: For the *iris* dataset, one observation worth mentioning is that the first row and first column of the synthetically generated data in the heatmap (Fig. 12) as well as the distributions in Fig. 5 did not correspond well with their real counterparts. This is because one of the features in the first row and the first column is the "Id" field, which served as the primary key. Its value has no intrinsic worth and is just used to distinguish each record. Hence, its correlation is not of much value. On a similar note, for the breast cancer dataset, correlation heatmaps were obtained as shown in Fig. 13 and Fig. 14 which show that the correlation of continuous columns for *adult* dataset is effectively captured by CTGAN [12] while from Fig. 15, observations were made that the CTGAN [12] extrapolated correlations for the credit dataset.

The analysis of correlation heatmaps for various datasets clearly shows that the complex relationships among the data features are effectively captured by the generative model CTGAN [12] when training is fine-tuned with optimum hyperparameters for each dataset.

TABLE IV. IRIS

Metric	Batch Size	Epochs	TVAE	CTGAN
CS Test	20	100	0.989522	0.999467
		500	0.997869	0.997204
		5000	0.996406	0.985703
	50	100	0.968894	0.992429
		500	0.985703	0.979415
		5000	**0.997**	**0.997**
	100	100	0.994283	0.999067
		500	0.986755	0.992429
		5000	0.983603	0.996274
KS Test	20	100	0.86533	0.706667
		500	0.941333	0.88000
		5000	0.950667	0.90000
	50	100	0.829333	0.714667
		500	0.896000	0.852000
		5000	**0.949**	**0.908**
	100	100	0.878667	0.730667
		500	0.905333	0.842667
		5000	0.910667	0.88000
KL_c	20	100	0.251850	0.113536
		500	0.307945	0.185478
		5000	0.384911	0.258259
	50	100	0.227984	0.114429
		500	0.311331	0.171203
		5000	**0.375**	**0.266**
	100	100	0.282140	0.118934
		500	0.262064	0.140582
		5000	0.338058	0.209723

TABLE V. BREAST CANCER

Metric	Batch Size	Epochs	TVAE	CTGAN
CS Test	50	100	0.953622	0.991299
		500	0.953622	0.947832
		5000	0.944939	**0.994**
	100	100	0.852929	0.976801
		500	0.979700	0.976801
		5000	0.997100	0.988399
KS Test	50	100	0.844492	0.671183
		500	0.887239	0.706729
		5000	0.881966	**0.863**
	100	100	0.821475	0.666194
		500	0.894042	0.719315
		5000	0.884914	0.850728
KL_c	50	100	0.697478	0.364236
		500	0.694817	0.389121
		5000	0.700066	**0.560**
	100	100	0.710529	0.365040
		500	0.706662	0.359388
		5000	0.687059	0.544663

VI. CONCLUSION

The paper presents a comprehensive overview of synthetic data generation and evaluation techniques and performs a

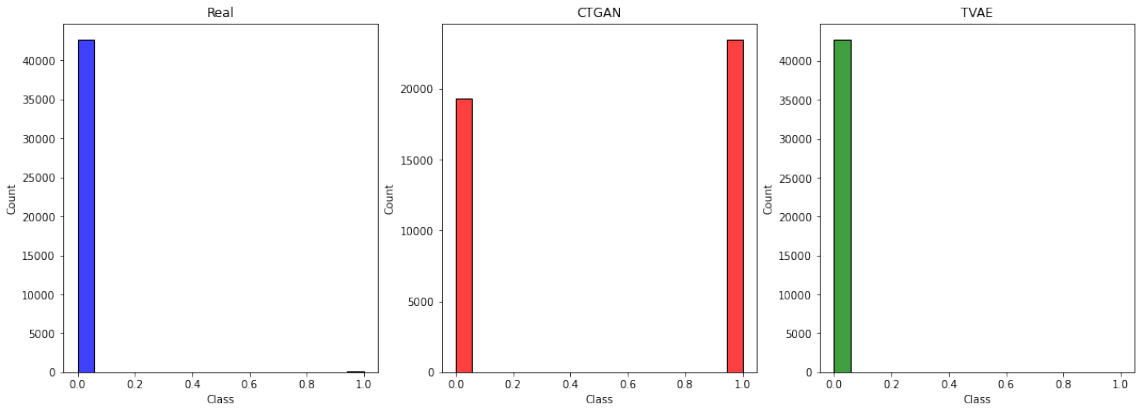

Fig. 11. Credit class imbalance.

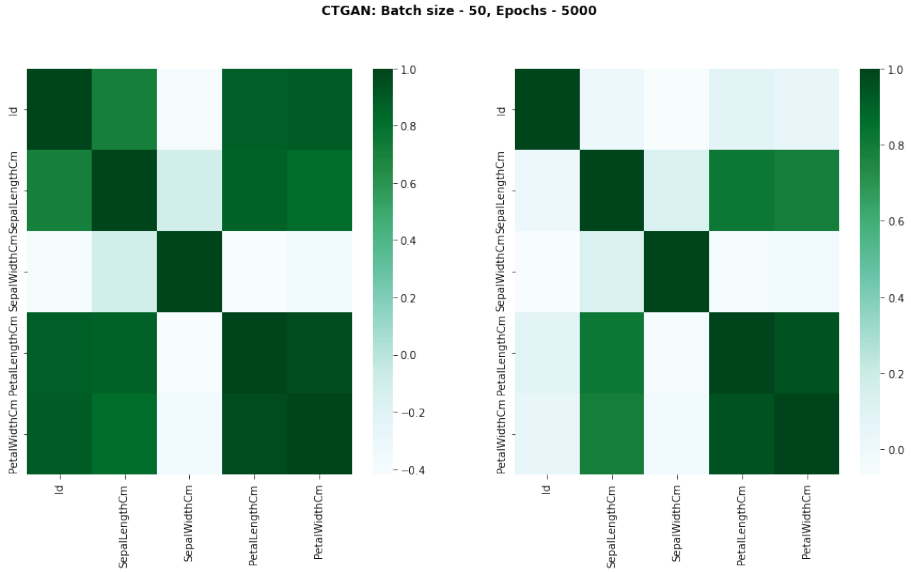

Fig. 12. Correlation Heatmap of Real (LHS) vs Synthetic (CTGAN)(RHS): Iris.

TABLE VI. ADULT

Metric	Batch Size	Epochs	TVAE	CTGAN
CS Test	300	200	0.986172	0.982407
		300	**0.990**	**0.982**
KS Test	300	200	0.845557	0.891655
		300	**0.855**	**0.886**
KL_c	300	200	0.866789	0.917393
		300	**0.935**	**0.935**
KL_d	300	200	0.942063	0.860879
		300	**0.951**	**0.858**

TABLE VII. CREDIT

Metric	Batch Size	Epochs	TVAE	CTGAN
KS Test	300	200	0.684782	0.656325
KL_c	300	200	0.854645	0.632121

TABLE VIII. IMPLEMENTATION ENVIRONMENT

Language	Python (version 3.11.0)
Tool	VS Code, Google Colaboratory
Libraries	Pandas, NumPy, Scikit Learn, Matplotlib, Seaborn, SciPy and SDV

rigorous analysis of small, medium, and large-scale synthetic data generated using two state-of-the-art generative models, TVAE and CTGAN. The choice of hyperparameters greatly influenced the quality of synthetic data. Small datasets (*iris* and *breast cancer*) required longer training periods for generating statistically similar synthetic data. Preserving univariate and bivariate distributions as shown in Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9 and Fig. 10 and multivariate joint distributions as shown in Fig. 12, Fig. 13 and Fig. 14 are achieved for small (*iris* and *breast cancer*), medium (*adult*) and large (*credit*) datasets using generative models. There is scope in the future to train large imbalanced datasets more rigorously and for more iterations with different parameters on high-end computational

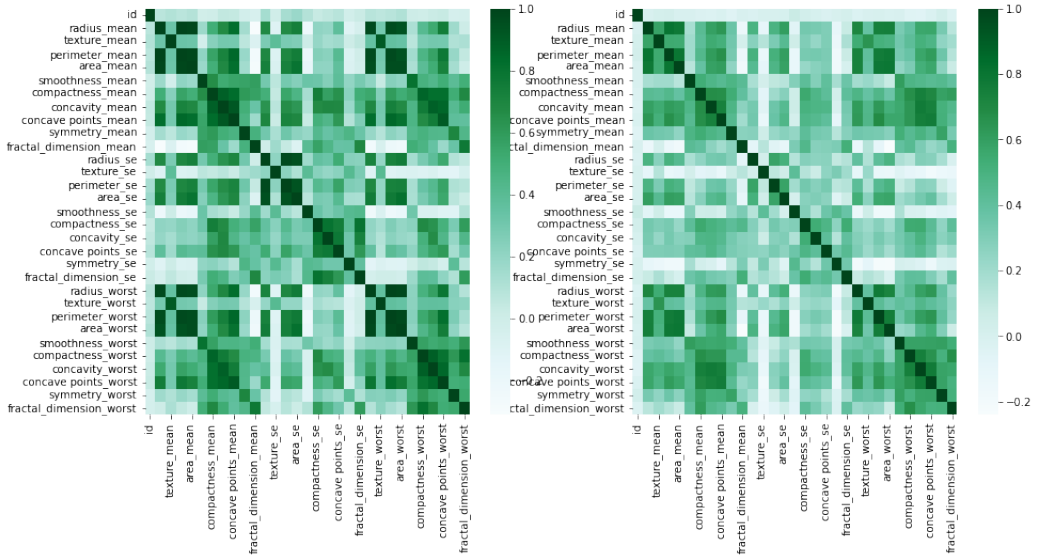

Fig. 13. Correlation Heatmap: Real (LHS) vs Synthetic (CTGAN)(RHS) - Breast Cancer.

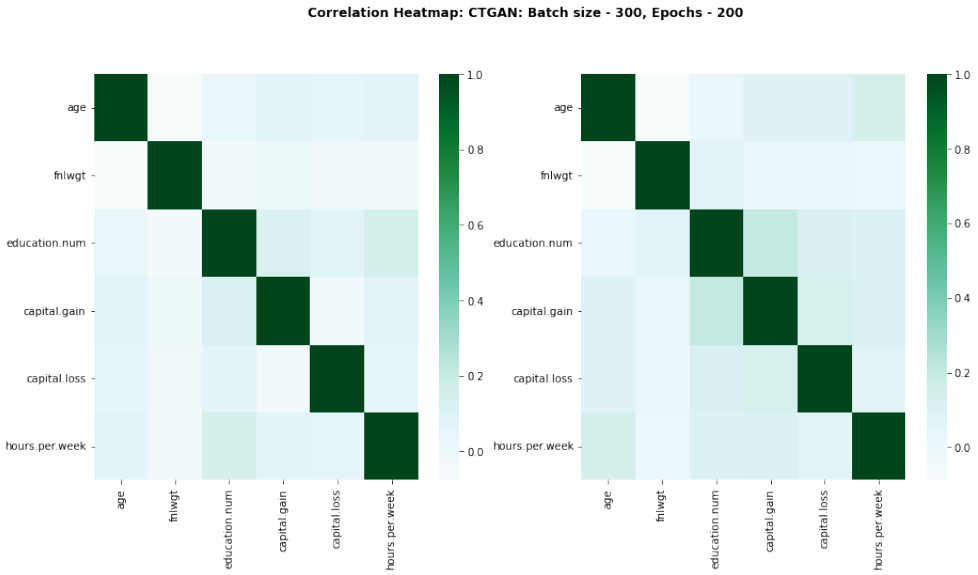

Fig. 14. Correlation Heatmap of Real (LHS) vs Synthetic (CTGAN)(RHS) data: Adult.

systems. While results for TVAE outperformed CTGAN for all four datasets by varying margins, as reflected by the KL Divergence score, CTGAN is the preferred method for generating privacy-preserving synthetic data due to its agnostic nature to real data values.

In this paper, the data on statistical metrics were evaluated, and a visualization report was presented to extensively analyze the synthetic data. The results not only highlighted the quality of synthetic data but also mentioned the shortcomings and caveats in the existing methods, which would open further dimensions in the line of research.

REFERENCES

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM, 63(11), pp.139-144.

[2] Dwork, C., 2008, April. Differential privacy: A survey of results. In International conference on theory and applications of models of computation (pp. 1-19). Springer, Berlin, Heidelberg.

[3] Mahoney, M.W., 2011. Randomized algorithms for matrices and data. Foundations and Trends® in Machine Learning, 3(2), pp.123-224.

[4] Muralidhar, K., Parsa, R. and Sarathy, R., 1999. A general additive data perturbation method for database security. management science, 45(10), pp.1399-1415.

[5] Žežula, I., 2009. On multivariate Gaussian copulas. Journal of Statistical Planning and Inference, 139(11), pp.3942-3946.

[6] Benali, F., Bodénès, D., Labroche, N. and de Runz, C., 2021. Mtcopula: Synthetic complex data generation using copula. In 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) (pp. 51-60).

[7] Zeng, Z. and Wang, T., 2022. Neural Copula: A unified framework for estimating generic high-dimensional Copula functions. arXiv preprint arXiv:2205.15031.

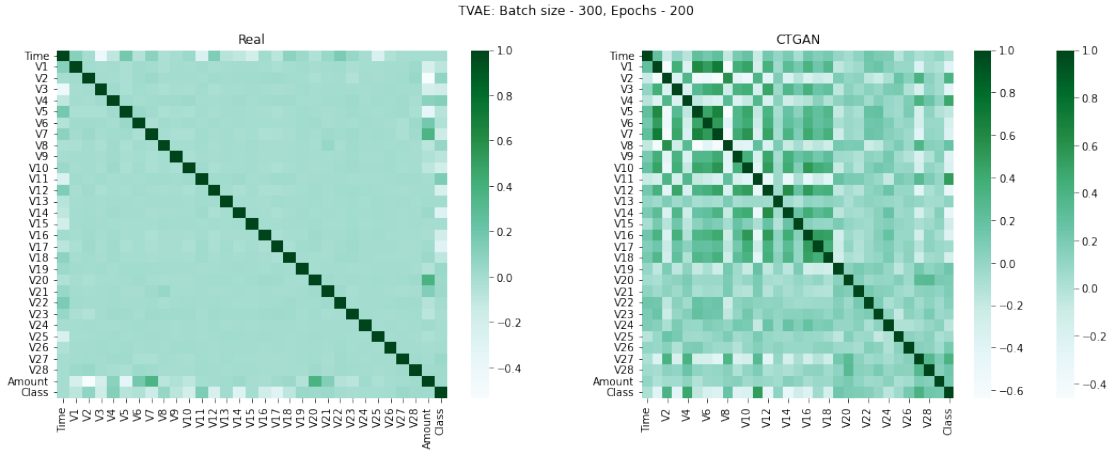

Fig. 15. Correlation Heatmap of Real (LHS) vs Synthetic (CTGAN)(RHS) data: Credit.

[8] Xu, L. and Veeramachaneni, K., 2018. Synthesizing tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264.

[9] Sun, Y., Cuesta-Infante, A. and Veeramachaneni, K., 2019, July. Learning vine copula models for synthetic data generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 5049-5057).

[10] Zhao, Z., Kunar, A., Birke, R. and Chen, L.Y., 2021, November. Ctab-gan: Effective table data synthesizing. In Asian Conference on Machine Learning (pp. 97-112). PMLR.

[11] Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. The annals of mathematical statistics, 22(1), pp.79-86.

[12] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems, 32.

[13] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

[14] Reiter, J.P., Wang, Q. and Zhang, B., 2014. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. Journal of Privacy and Confidentiality, 6(1).

[15] Sklar, A., 1973. Random variables, joint distribution functions, and copulas. Kybernetika, 9(6), pp.449-460.

[16] Demarta, S. and McNeil, A.J., 2005. The t copula and related copulas. International statistical review, 73(1), pp.111-129.

[17] Czado, C., 2019. Analyzing dependent data with vine copulas. Lecture Notes in Statistics, Springer, 222.

[18] Cavanaugh, J.E. and Neath, A.A., 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), p.e1460.

[19] Yu, Y., Si, X., Hu, C. and Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), pp.1235-1270.

[20] Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4), pp.307-392.

[21] Ben-Gal, I., 2008. Bayesian networks. Encyclopedia of statistics in quality and reliability, 1.

[22] Gogoshin, G., Branciamore, S. and Rodin, A.S., 2021. Synthetic data generation with probabilistic Bayesian Networks. Mathematical biosciences and engineering: MBE, 18(6), p.8603.

[23] Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B. and Sweeney, L., 2019. Privacy preserving synthetic data release using deep learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 510-526). Springer, Cham.

[24] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[25] Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S. and Yang, B., 2020. MedGAN: Medical image translation using GANs. Computerized medical imaging and graphics, 79, p.101684.

[26] Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H. and Kim, Y., 2018. Data synthesis based on generative adversarial networks. arXiv preprint arXiv:1806.03384.

[27] Arjovsky, M., Chintala, S. and Bottou, L., 2017, July. Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.

[28] Gauthier, J., 2014. Conditional generative adversarial nets for convolutional face generation. Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester, 2014(5), p.2.

[29] Bishop, C.M. and Nasrabadi, N.M., 2006. Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[30] Zhao, Z., Kunar, A., Birke, R. and Chen, L.Y., 2022. CTAB-GAN+: Enhancing Tabular Data Synthesis. arXiv preprint arXiv:2204.00401.

[31] Lederrey, G., Hillel, T. and Bierlaire, M., 2022. DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data. arXiv preprint arXiv:2203.03489.

[32] Jordon, J., Yoon, J. and Van Der Schaar, M., 2018, September. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International conference on learning representations.

[33] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. and Erlingsson, Ú., 2018. Scalable private learning with pate. arXiv preprint arXiv:1802.08908.

[34] Xie, L., Lin, K., Wang, S., Wang, F. and Zhou, J., 2018. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739.

[35] Massey Jr, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 46(253), pp.68-78.

[36] McHugh, M.L., 2013. The chi-square test of independence. Biochemia medica, 23(2), pp.143-149.

[37] Villani, C., 2009. Optimal transport: old and new (Vol. 338, p. 23). Berlin: springer.

[38] Hoofnagle, C.J., van der Sloot, B. and Borgesius, F.Z., 2019. The European Union general data protection regulation: what it is and what it means. Information & Communications Technology Law, 28(1), pp.65-98.

[39] Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), pp.91-110.