International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00653-3

REGULAR PAPER

A comparative exploration of two diffusion generative models on tabular data synthesis

Neetu Kumari[1] · **Enayat Rajabi**[1]

Received: 23 June 2024 / Accepted: 20 September 2024

Abstract
The generation of synthetic tabular data has become increasingly important as a solution to data accessibility, privacy, and resource constraints issues. Despite the development of various synthetic data generators, comprehensive evaluations of their effectiveness and generalizability are limited. This study aims to conduct a detailed comparative analysis of two notable diffusion models, TabDDPM and TabSyn, which are reputed for their ability to produce high-quality synthetic data but have not been extensively tested against each other in the context of tabular datasets. The research proposes a framework to evaluate the performance of TabSyn and TabDDPM models in terms of generating synthetic tabular data, maintaining statistical similarity with real datasets, and upholding data privacy. We utilized six datasets with varying dimensionality and assessed the generated data using the proposed framework including three key criteria: similarity, utility, and privacy preservation. Our findings indicate that TabSyn surpasses TabDDPM in similarity, privacy, and utility metrics for all datasets. The quality of data produced by the TabSyn model closely mirrors real datasets, showing strong statistical alignment for both categorical and continuous variables. Additionally, the utility of the synthetic data for machine learning applications is comparable to that of real tabular data. Privacy assessments confirm that the data generated by the TabSyn model maintain stringent privacy standards. The findings confirm TabSyn's effectiveness and establish its superiority in generating synthetic data, providing guidance for practitioners in selecting synthetic data generation models and setting benchmarks for future innovations in the field. The results underscore the importance of adopting advanced generative models like TabSyn to create tabular data that not only accurately reflect real-world distributions but also protect individual privacy. Code has been made available at GitHub.

Keywords Generative models · Synthetic data · Diffusion models · TabDDPM · TabSyn · Privacy preservation · Tabular data · Tabular healthcare data

1 Introduction

In the modern digital environment, people engage in various activities such as browsing Google, utilizing social media platforms, conducting online transactions, and engaging in e-commerce, which generates abundant data. Accumulating a substantial volume of data is an important resource for different sectors, aiding decision-making and supporting research efforts [1]. Researchers encounter difficulties accessing real-world data in specific scenarios, such as investigating hypothetical situations or exploring future instances where current data are unavailable—such as during the COVID-19 pandemic [2]. Nevertheless, constraints on accessing authentic data arise from privacy concerns and legal regulations within sensitive domains like healthcare and finance [1].

The foundation of modern artificial intelligence (AI) lies in data. Acquiring the appropriate data is crucial and presents a significant challenge when developing robust AI models. Gathering high-quality, real-world data are intricate, costly, and time intensive. Synthetic data have arisen as a solution to this issue. Due to its versatility and usefulness, its inception led to widespread adoption across various sectors, such as finance and healthcare. Synthetic data refer to artificially

Neetu Kumari and Enayat Rajabi have contributed equally to this work.

✉ Neetu Kumari
neetunsca@gmail.com

Enayat Rajabi
enayat_rajabi@cbu.ca

[1] Department of Management Science, Cape Breton University, Sydney, NS, Canada

Published online: 05 October 2024

generated data replicating real-world data's statistical properties and characteristics [3]. It is generated by algorithms and mathematical models, often based on machine learning techniques, to mimic the patterns and distributions found in real data. Utilizing synthetic data allow researchers to overcome the limitations associated with accessing sensitive or unavailable real data, as well as the time and expense required to obtain such data. This method also facilitates the conduct of what-if analyses and the simulation of hypothetical scenarios [3].

Synthetic data generation has become increasingly popular in the last few decades for all types of data generation, such as images, sounds, text to images, or tabular data with varying distributions and data types. Various models have been developed to generate artificial data, such as GAN [4]-based, VAE [5]-based, and diffusion models. Using synthetic data in sectors that handle sensitive information, such as healthcare and finance, presents distinct challenges and advantages. The benefits of using synthetic data are substantial; it mitigates the risk of exposing sensitive details when sharing data for advancement, research, and innovation and reduces the time needed to access such data. Additionally, synthetic data facilitate the creation of diverse and customized datasets that can be utilized for various purposes, including simulations, predictive research, hypothesis testing, and model validation. However, the use of synthetic data also involves significant risks. One major concern is the risk of re-identification, where synthetic data, if not adequately anonymized, can potentially be traced back to the original data points. Moreover, generating high-quality synthetic data pose challenges; the quality of synthetic data is affected by the volume of entries, the data's complexity, and the specific generative model employed in its creation. These factors must be carefully managed to maximize the benefits of synthetic data while minimizing potential risks [6].

Numerous models [7–9] have been developed to address these challenges over time, each aiming to improve the quality of artificial data. These models are continuously evaluated and updated based on various metrics to enhance the generated data's quality. In healthcare, generative models have been designed to replicate diverse medical data types, including radiology and pathology images, electronic health records, genomics data, and clinical trials. Notable examples include medGAN [10], EMR-WGAN [11], ADS-GAN [4], CTABGAN [12], SparseGAN [7], and CTAB-GAN+ [13]. In the current landscape, diffusion probabilistic models have emerged as the primary paradigm for generative modeling across essential data modalities. Widely embraced in the computer vision community, these models have demonstrated substantial advancements compared to GANs. Examples showcasing the superiority of diffusion models in specific tasks include GLIDE [14] for text-to-image generation and Diffwave [15] for audio synthesis. The progression and modification of models continue, with the latest high-performing diffusion models including EHRDiff [16], MedDiff [17], TabDDPM [18], and TabSyn [8]. These models aim to overcome the challenges associated with generating tabular data, particularly in maintaining the integrity and relevance of the generated data [19, 20].

Research [9, 21] in synthetic data is progressing in two main areas: (i) some researchers are developing new methods for generating synthetic data, while (ii) others are examining the effectiveness of these generators in practical situations. Despite the development of high-performing diffusion generative models such as TabDDPM and TabSyn, there is a notable lack of comparative studies on these models for tabular datasets. This study compares these two models and evaluates the generated data on various metrics to identify the most suitable model for creating tabular synthetic data that closely mirror real data in terms of statistical distributions while also ensuring privacy protection. To accomplish this, we propose a comprehensive evaluation framework encompassing based on similarity, utility, and privacy preservation. The metrics used include the Kolmogorov–Smirnov distance, total variation test, and tests for differential privacy measures like α-precision and β-recall.

Figure 1 illustrates a flowchart depicting the synthetic data generation process using TabSyn and TabDDPM diffusion models, then evaluating the generated data using various metrics to determine the optimal generative model.

2 Related work

Table 1 presents descriptions of the 7 selected publications and conducts a comparative analysis between our study and these publications. The table delineates the models employed, the types of variables considered across various datasets, and the diverse evaluation methods utilized in these studies.

Advanced techniques such as GAN models [4, 4, 8, 22–25], VAE models [5, 8, 23], and diffusion models [18, 18, 20, 26, 27] have been instrumental in generating highly realistic artificial tabular data. The efficacy of generated tabular data has been assessed using a broad spectrum of metrics that evaluate similarity, utility, and privacy. Notably, studies in references [8, 18] and [28] have performed comparative analyses among these generative methods, often showing superior performance by diffusion models, i.e., TabSyn and TabDDPM. Although many studies focus on synthesizing mainly categorical, numerical, and time-series data, the evaluation metrics demonstrate significant diversity. There is a notable lack of uniformity in evaluation approaches, with not all studies assessing the three defined dimensions of resemblance, utility, and privacy. Often, only one or two of these dimensions are evaluated, and the metrics used are limited,

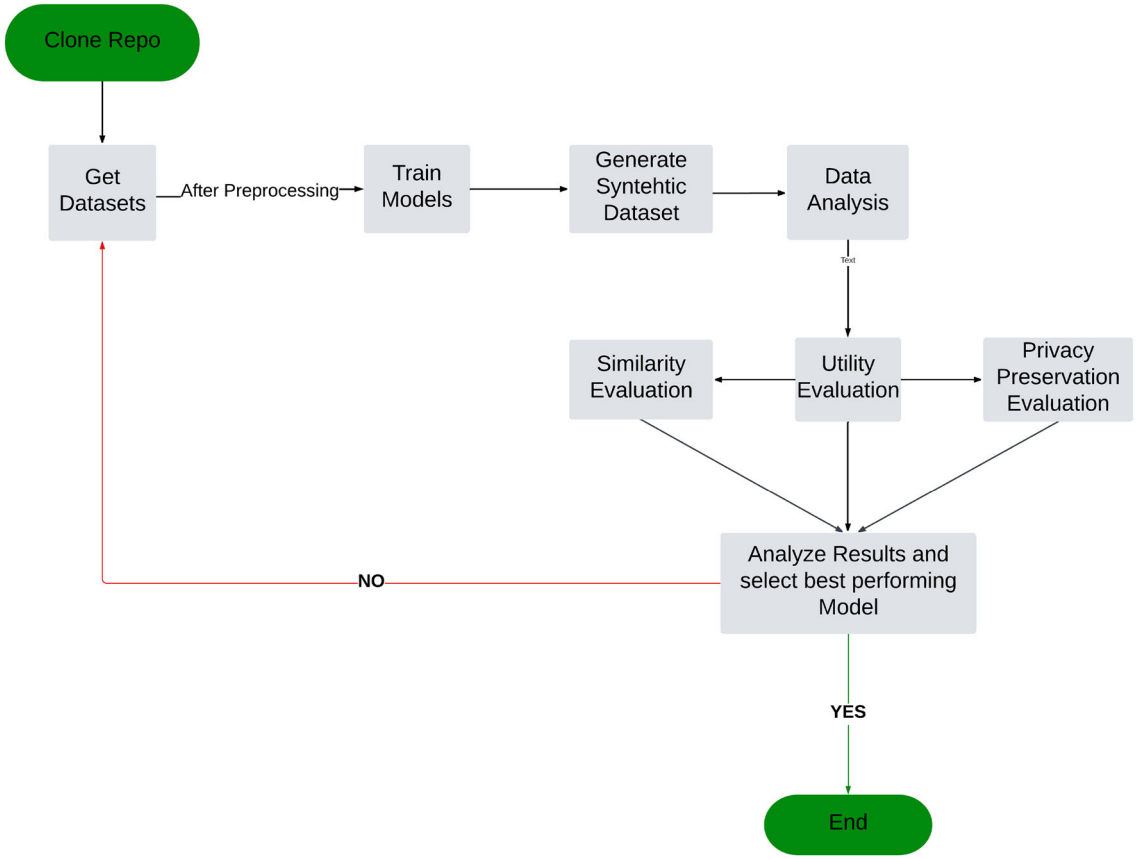


Fig. 1 Flowchart illustrates the steps from data collection to model evaluation, where the synthetic dataset is assessed based on three key evaluation metrics: similarity, utility, and privacy preservation. The decision on the best performing model is made by considering the combined results of these three metrics equally, ensuring a holistic evaluation of model performance

particularly for privacy, which is the least evaluated dimension in the reviewed publications.

In response to these observations, we reviewed various models and ultimately selected TabSyn and TabDDPM for their innovative application of diffusion models. This decision was also influenced by the noticeable scarcity of comparative studies specifically focusing on these two models. Some studies including [18] and [28] evaluated TabDDPM against well-known models like TVAE, CTABGAN, demonstrating TabDDPM's superior performance in generating tabular data, whereas [8] indicated that TabSyn often surpasses TabDDPM. However, these studies did not explore a broad spectrum of datasets with varying sizes and complexities.

To bridge this gap, we conducted a detailed comparative analysis of TabDDPM and TabSyn across a diverse range of datasets, including both numerical and categorical variables. We used a comprehensive set of methods to assess the similarities and differences in the data generated by each model.

We adopted a variety of evaluation metrics to assess the generated data across three main criteria: similarity, utility, and privacy. We meticulously selected these metrics from existing literature [4, 23, 29] and [8] to provide a holistic assessment of the synthetic data across these key dimensions. The similarity evaluation was subdivided into Variable Correlation, Distribution Similarity, and Pair-wise Correlation, utilizing various metrics to clarify how closely the artificial data resemble real data for different data types. The utility was evaluated using the Training on Synthetic and Testing on Real(TSTR) [8] approach, which is widely used in the literature. For privacy preservation, we combined various metrics such as alpha-precision, beta-recall, and distance to closest record (DCR) to determine which model performs better in terms of privacy, a critical factor in deciding the overall efficacy of a model.

3 Background and fundamentals

Decades of research have transformed the landscape of synthetic data generation. It began with early algorithms like SMOTE (synthetic minority oversampling techniques) for

Table 1 Comparative table of previous publication

Publications	DataType	Methods	Evaluation methods
Arjovsky [29]	Images	WGAN	Total Variation distance
			Kullback–Leibler divergence
			Jensen–Shannon divergence
			Earth mover distance or Wasserstein-1
Dash [22]	Numerical	timeGAN	Compare average trends
	Categorical	healthGAN	Welsch t-test
	Binary		ML models: TRTR, TRTS, TSTS and TSTR
	Time-series		
Rashidian [24]	Numerical	cGAN	MAE for means and sd
	Categorical	AC-GANWGAN	Compare Pearson correlation coefficients
	Binary	WGAN-GP	ML models: TRTR, and TSTR
		SmoothGAN	MMD
Yoon [4]	Numerical	ADS-GAN	Student t-test and Chi-squared test
	Categorical	PATE-GAN	ML models: TSTR
	Binary	DP-GAN	JSD and Wasserstein distance
		medGAN	
		WGAN-GP	
Lee [23]	Categorical	DAAE	ML model to classify records
	Time-series	medGAN	clinical experts
		VAE	DBSCAN
		VAE-GAN	ML models: TSTR
		WAE	Differential privacy cost
		ARAE	
Wang [25]	Numerical	DP-GAN	Statistics and cumulative distributions
	Time-series	PART-GAN	Inception Score
			Euclidean distances
Zhang [8]	Numerical	CTGAN	Low order statistics: Column-wise density
	Categorical	TVAE	Pair-wise column correlation
		GOOGLE	High-order metrics: alpha-precision, beta-recall
		GReaT	Machine learning efficiency (MLE)
		STaSy	
		CoDi	
		TabDDPM	
		TabSyn	
Our study	Numerical	TabDDPM	Statistical Methods
	Categorical	TabSyn	Machine learning efficiency: TSTR
			Privacy preservation: α-precision, β-recall, DCR

generating artificial data. Over time, the concept of synthetic data evolved, leading to the proposal of various variants [30]. The real breakthroughs, however, occurred with the rise of deep learning. Innovative approaches such as VAE, GAN, and later Diffusion models emerged, offering more promising and advanced outcomes in synthetic data generation [31]. Advancements in GANs and Diffusion models have successfully addressed challenges posed by earlier models, particularly in generating tabular data with mixed types, non-Gaussian distribution, multi-modal distribution, and highly imbalanced datasets. These developments not only overcome previous limitations but also enhance the quality of synthetic data. The generated data now closely resemble real data and preserve sensitivity. Continuous innovation and upgrades in generative models contribute to the improved performance of synthetic data. Over time, extensive modifications in evaluation metrics have played a crucial role in refining these model types [32].

3.1 GAN-based models

Generative adversarial network (GAN) [4, 10–12] and [13] formulates the generation problem into a supervised learning task. Specifically, it comprises two neural network modules: discriminator and generator. The objective of the generator is to generate data that are close to the real data. On the other hand, the objective of the discriminator is to discriminate the fake data (generated by the generator) from the real ones. It performs a binary classification task, where the real data from the training set are regarded as the positive samples; the generated data (by generator) are regarded as the negative samples; generator and discriminator are trained in a mini-max manner. Over the time, various GAN models have been developed utilized for image generation, audio video generation, text, and tabular data generation.

3.2 Diffusion-based models

Diffusion models, initially described by [26] and further developed by [20], are a type of generative modeling designed to approximate a target distribution at the end of a Markov chain starting from a standard Gaussian distribution. Each step of this Markov chain involves a deep neural network that is adept at reversing the diffusion process using a Gaussian kernel. These models have been shown to gradually transform a simple known distribution into a target distribution through an iterative denoising process, a concept that parallels score matching, as illustrated by [20] and further discussed by [26].

Recent innovations by [27] and [33] have introduced more sophisticated model architectures and advanced learning protocols, which have propelled diffusion models to exceed GANs in generative quality and diversity. In this research, we adapt diffusion models TabSyn and TabDDPM to address tabular data challenges.

3.2.1 TabDDPM

TabDDPM is a diffusion model that can generate synthetic tabular data with mixed data types, such as numerical and categorical features. TabDDPM uses multinomial diffusion for categorical features and Gaussian diffusion for numerical features. TabDDPM can handle any tabular dataset and any feature types, and it outperforms existing GAN/VAE alternatives on a wide set of benchmarks. TabDDPM can also provide privacy-preserving synthetic data that do not reveal the original records [18].

According to [18], TabDDPM has the following limitations:

1. It overlooks temporal or spatial dependencies among tabular records, potentially crucial for certain applications. A potential enhancement could involve incorporating spatio-temporal diffusion models.
2. It does not differentiate between various numerical features, such as real valued, positive real valued, or ordinal. A plausible extension could involve employing distinct diffusion processes for different numerical types.

3.2.2 TabSyn

The TabSyn model, designed for mixed-type tabular data, comprises a variational autoencoder (VAE) and a score-based diffusion model. The VAE transforms the original tabular data into a continuous latent space, where each latent variable corresponds to a data column. Meanwhile, the score-based diffusion model generates synthetic data from the latent space through reverse steps, introducing and removing Gaussian noise. This process is guided by a score function assessing the data's likelihood. The TabSyn model excels in managing diverse data types, capturing inter-column relationships, and generating synthetic data of high quality [8].

In the context of [8], the TabSyn model presents certain limitations:

1. Data preprocessing: TabSyn necessitates preprocessing steps, including handling missing values and transforming numerical and categorical features. This process might introduce noise or bias to the data.
2. Latent space dimensionality: The model employs a fixed dimensionality for the latent space, which may not be optimal for various datasets or tasks. An enhancement could involve implementing an adaptive latent space capable of adjusting its dimensionality based on data complexity.
3. Generalization to other data types: While TabSyn is crafted for mixed-type tabular data, it may encounter challenges handling different data types like text, images, or graphs. A prospective avenue involves exploring ways to extend TabSyn to diverse domains and modalities.

4 Data collection and preprocessing

We have selected six real-world datasets, each containing both numerical and categorical attributes. These datasets are sourced from the UCI Machine Learning Repository[1] and Kaggle,[2] including Diabetes, Adult, Insurance, Asthma, Heart, and Obesity. The reasons for choosing these datasets are their variability in feature data types, diversity in entry volumes, and the lack of comparative studies between the TabSyn and TabDDPM models for healthcare tabular

[1] https://archive.ics.uci.edu/datasets

[2] https://www.kaggle.com/datasets

Table 2 Statistics of the datasets used in the study

Dataset	# Entries	# Numerical	#Categorical	# Train	#Test	Task
Obesity	2111	8	9	1899	212	Categorical MultiClass
Diabetes	253680	7	15	228312	25368	Categorical BiClass
Adult	48842	6	8	32561	16281	Categorical BiClass
Insurance	1338	2	4	1204	134	Regression
Heart	5000	5	7	4500	500	Categorical BiClass
Asthma	300	2	3	270	30	Categorical BiClass

datasets in the existing literature. Our one of the objectives was to identify the best synthetic data generator for datasets of different sizes, such as small datasets (datasets that we used in our study with entries of less than 3,000 and varying numbers of features) and large datasets (datasets used in the study ranging from 48,000 to approximately 250,000 entries) [34–36].

Table 2 presents the statistics of the datasets used in our study. The number of rows (# Rows), numerical features (# Num), and categorical features (# Cat) are listed for each dataset. The target column is classified as either numerical or categorical based on the task type. Specifically, the target column is considered categorical for classification tasks, whereas for other tasks, it is treated as numerical. Each dataset is divided into training, validation, and testing sets to facilitate the Machine Learning Efficiency experiments.

Below is a detailed introduction to each dataset:

1. **Diabetes**[3]: This dataset, known as the Diabetes Health Indicators Dataset, comprises healthcare statistics and lifestyle survey information, along with diagnoses of diabetes. The target variable for classification, "Diabetes_binary," indicates whether a patient has diabetes, is pre-diabetic, or is healthy.
2. **Adult**[4]: The "Adult Census Income" dataset contains demographic and employment-related features. The task is to predict whether an individual's income exceeds 50,000. It includes a categorical target variable "income" with two classes.
3. **Obesity**[5]: The dataset contains data for estimating obesity levels in individuals from Mexico, Peru, and Colombia, based on their dietary habits and physical condition. It includes a categorical target variable labeled "NObesity," with categories such as "Sufficient Weight," "Normal Weight," "Overweight Level I," "Overweight Level II," "Obesity Type I," "Obesity Type II," and "Obesity Type III."
4. **Heart**[6]: This dataset contains records of 5000 patients, each with 13 clinical features including age, diabetes, blood pressure, and more. The target variable is a classification feature called "death_event," which indicates whether a patient died during the follow-up period.
5. **Insurance**[7]: This dataset contains information on 1338 individuals 7 personal attributes such as age, gender, BMI, family size, and more. The target variable is the medical insurance cost. This dataset can predict an individual's medical insurance cost based on their personal and geographic features.
6. **Asthma**[8]: The asthma dataset includes data on 300 individuals, each with 6 personal attributes such as age, smoking status, and more. The goal is to predict whether an individual has asthma based on these attributes.

For our study, we utilize the source code of TabSyn and TabDDPM obtained from the GitHub Repository referenced in the paper [8]. We cloned the repository to our local system, added the datasets to the required directory, and made necessary modifications to the source code. Following data preprocessing, we partitioned all the datasets into training and testing subsets, allocating 90% of the data for training and 10% for testing.

5 Synthetic data generation

This study used TabDDPM (Tabular Denoising Diffusion Probabilistic Model) and TabSyn to generate tabular synthetic data.

In our study, we trained TabDDPM for 1000 epochs and the VAE for 1000 epochs. Following the VAE training, we proceeded to train the Diffusion Model for an additional 1001 epochs. The NVIDIA GeForce RTX 4050 Laptop GPU was utilized to expedite the training process. Notably, the VAE

[3] https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[4] https://archive.ics.uci.edu/dataset/2/adult

[5] https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.zip

[6] https://www.kaggle.com/datasets/aadarshvelu/heart-failure-prediction-clinical-records

[7] https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance

[8] https://www.kaggle.com/datasets/jatinthakur706/copd-asthma-patient-dataset

model training was the most time-consuming, taking approximately five hours for the diabetes dataset and roughly four hours for the obesity dataset.

Following training, we generated synthetic datasets of equal size to the real data using both trained models. The generation of the artificial dataset for the smaller datasets (Asthma, Obesity, Heart, Insurance) was swift, especially with the TabDDPM model. However, generating synthetic datasets for the larger datasets (Diabetes and Adult) was time-consuming, particularly for the TabSyn model.

6 Results and discussion

The study underscores the importance of protecting privacy and maintaining data utility in tabular datasets. It employs a range of evaluation metrics to assess these aspects, including:

i) Similarity evaluation: Variable correlation: Statistical similarity of each variable is assessed by computing the mean and standard deviation for continuous variables and analyzing the percentage of each category for categorical variables. Distribution similarity: Kolmogorov–Smirnov Test, Chi-square test, Wasserstein Distance, and Overall Density estimation of single columns are utilized to measure distribution similarity. Pair-wise Correlation: Pair-wise correlation between variables is examined.
ii) Utility Evaluation: Usability of synthetic data in Machine Learning tasks is evaluated to assess its utility.
iii) Privacy Preservation Measures: Distance to closest record (DCR) is calculated to evaluate privacy preservation. Alpha-precision and beta-recall are also computed as privacy preservation measures.

6.1 Similarity evaluation

A comprehensive analysis of their similarities is conducted using various evaluation metrics to evaluate the degree of similarity between synthetic and real data. These metrics provide insights into the effectiveness of the generated data and help identify the superior performance among TabDDPM and TabSyn.

Variable Correlation:

Statistical similarity of continuous variables is assessed, and the percentage of each category for categorical variables is analyzed.

Table 3 presents the outcomes of Variable Correlation for Continuous variables for two of the datasets Obesity and Diabetes. It offers a comparative view of the similarity between real and synthetic data generated by TabDDPM and TabSyn.

The results indicate that for the Obesity dataset for both continuous and categorical variables, TabSyn and TabDDPM demonstrate similar performance. Conversely, for the Diabetes dataset, TabSyn exhibits superior behavior over TabDDPM for both variables types. Additionally, experiments with other datasets revealed that TabSyn generally performs better across all types of datasets.

Distribution Similarity:

To assess the similarity of distributions, we conducted visual comparisons (in Fig 2a and b) by plotting the distributions of both real and synthetic data. This analysis included comparing synthetic data generated by TabDDPM and TabSyn with the Obesity and Diabetes datasets, covering both continuous and binary variables (Fig 2c and d). To obtain statistical insights into the distribution of variables, we employed the Kolmogorov–Smirnov Test for continuous variables and the Chi-square test for categorical variables.

Table 4 presents the results of the KS test for two datasets. For the Obesity dataset, the effectiveness of the generated synthetic data is inconclusive. However, for the Diabetes dataset, the synthetic data generated by TabSyn closely resemble the real data. Additional testing on other datasets indicates that TabSyn consistently outperforms TabDDPM across all datasets. Furthermore, higher p-values for the Chi-square test across almost all categorical variables in all the datasets suggest a lack of statistical significance. This indicates that differences in category distributions may be attributable to random chance rather than meaningful associations. Therefore, alternative tests are necessary to gain insights into categorical variables such as Total Variation Distance(TVD) and Contingency similarity. These metrics will provide a comprehensive view of both the distributional fidelity and the relational dynamics of the data. TVD helps assess the disparities in distribution density, while Contingency Similarity evaluates the strength of relationships between categorical variables.

Consequently, we opted to employ a more integrated metric to assess the similarity and differences between various types of variables in the datasets. We decided to evaluate column-wise distribution density estimation and pair-wise column correlation. We used the Kolmogorov–Smirnov Test (KST) for numerical columns and the Total Variation Distance (TVD) for categorical columns for density estimation. Pair-wise column correlation was assessed using Pearson correlation for numerical columns and contingency similarity for categorical columns. Performance was measured by comparing correlations computed from real and synthetic data. We initially categorize numerical values by grouping them into discrete intervals through bucketing to establish the correlation between numerical and categorical columns. Subsequently, we compute the corresponding contingency similarity, hence determined the overall quality of the dataset.

Table 3 Comparative table for real, synthetic data generated by TabDDPM and TabSyn for Continuous variables for both the obesity and diabetes datasets

Dataset/Metrics		Mean			Std		
		Real	TabDDPM	TabSyn	Real	TabDDPM	TabSyn
Obesity	Age	24.342	23.993	23.854	6.413	6.324	5.523
	Height	1.702	1.710	1.704	0.094	0.094	0.087
	Weight	86.498	87.241	85.793	26.310	26.659	25.369
	FCVC	2.418	2.404	2.436	0.532	0.542	0.514
	NCP	2.690	2.749	2.701	0.778	0.721	0.727
	CH20	2.005	2.026	1.995	0.615	0.582	0.569
	FAF	1.017	1.020	1.000	0.849	0.835	0.787
	TUE	0.655	0.679	0.704	0.611	0.591	0.587
Diabetes	BMI	28.382	62.188	28.669	6.610	42.389	6.761
	MentHlth	3.187	14.872	2.910	7.416	14.999	7.103
	PhysHlth	4.244	15.408	3.946	8.723	14.994	8.461

Table 4 Comparative table of KS Statistics and *p*-value for Real v/s TabDDPM and Real v/s TabSyn for Continuous variables for both the obesity and diabetes datasets

Dataset/Metric		Real vs TabDDPM KS stats, *p*-value	Real vs TabSyn KS stats, *p*-value
Obesity	Age	**0.039, 0.112**	0.046, 0.037
	Height	0.480, 0.025	0.046, 0.037
	Weight	0.044, 0.049	**0.031, 0.319**
	FCVC	**0.028, 0.450**	0.049, 0.021
	NCP	**0.073, 7.579**	0.048, 0.026
	CH20	0.055, 0.007	0.070, 0.0001
	FAF	0.045, 0.041	0.039, 0.103
	TUE	0.051, 0.014	0.066, 0.0005
Diabetes	BMI	0.583, 0.0	**0.022, 8.432**
	MentHlth	0.448, 0.0	**0.036, 8.421**
	PhysHlth	0.437, 0.0	**0.030, 3.150**

Bold values indicate the better output between the two models being compared

Table 5 Comparative table of overall quality score, column shapes, and column pair trends of generated data by **TabDDPM** for all datasets

Dataset/Metrices	Diabetes	Obesity	Adult	Insurance	Asthma	Heart
Overall quality score	62.25	**95.36**	51.31	36.73	31.93	43.61
Column shapes	72.04	96.37	58.27	47.56	47.10	54.12
Column pair trends	52.46	**94.35**	44.35	25.9	16.76	33.1

Bold values indicate the better output between the two models being compared

Tables 5 and 6 illustrate the challenge in determining the superior synthetic data for the obesity dataset, while it's evident that TabSyn outperforms for the diabetes dataset.

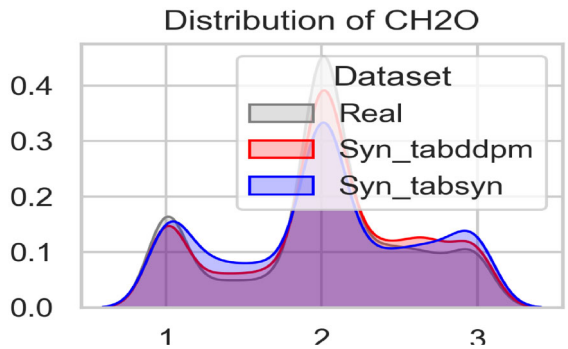
6.2 Utility evaluation

Real-world tabular datasets, particularly those in healthcare, often contain highly sensitive and personally identifiable information about individuals. As a result, their usage is severely restricted. Tabular data generation models aim to alleviate this concern by creating synthetic data that mimic

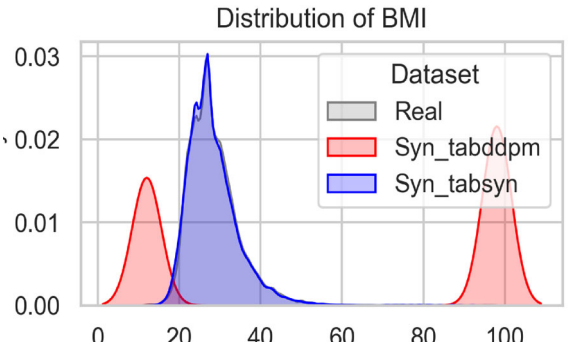
Table 6 Comparative table of overall quality score, column shapes, and column pair trends of generated data by **TabSyn** for all datasets

Dataset/Metrices	Diabetes	Obesity	Adult	Insurance	Asthma	Heart
Overall quality score	**97.94**	94.68	**98.15**	**95.51**	**90.44**	**80.87**
Column shapes	**98.51**	**96.84**	**98.66**	**97.41**	**93.40**	**90.13**
Column pair trends	**97.36**	92.52	**97.65**	**93.61**	**87.49**	**71.61**

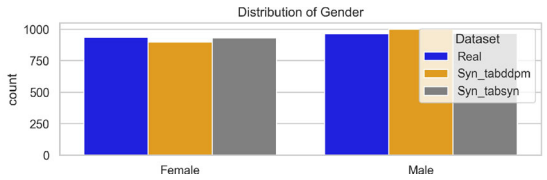
Bold values indicate the better output between the two models being compared



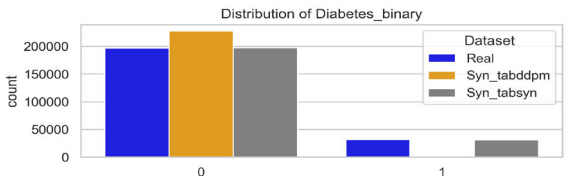
(a) **Numerical** attribute **CH2O** for the **Obesity** dataset



(b) **Numerical** attribute **BMI** for the **Diabetes** dataset



(c) **Categorical** attribute **Gender** for the **Obesity** dataset



(d) **Categorical** attribute **Diabetes_binary** for the **Diabetes** dataset

Fig. 2 Comparing the distribution of categorical and numerical attributes for the Diabetes and Obesity datasets with the generated datasets by TabDDPM and TabSyn

the characteristics of real data but cannot be linked back to the original dataset [6].

Machine Learning Efficiency (MLE) plays a vital role in safeguarding data privacy. It assesses whether a generative model can generate synthetic data that closely resembles real data by comparing the performance of a machine learning model trained on synthetic data and tested on real data, a process known as Training on Synthetic and Testing on Real (TSTR) test [8]. Additionally, MLE demonstrates the practical utility of synthetic data in real-world scenarios. Improved outcomes in MLE indicate that synthetic data can effectively replace real data for training machine learning models to make predictions and can also serve as a viable alternative for research purposes.

To evaluate the effectiveness of synthetic datasets, we partitioned the real data table into training and testing sets with a 9:1 ratio. The generative models were trained on the real training set, and a synthetic dataset of equivalent size was generated. These synthetic data were then utilized to train a classification or regression model, which was subsequently evaluated using the real testing set. The performance of MLE was quantified using the AUC score for classification tasks and RMSE for regression tasks.

Table 7 illustrates that for the smaller dataset, it is difficult to discern which synthetic data perform better in terms of MLE. However, for the larger dataset, it is clear that TabSyn generated data demonstrate superior utility for ML models.

6.3 Privacy preservation measures

Ensuring statistical similarity between real and synthetic data is crucial, but the preservation of sensitive information is equally important. Assessing privacy preservation is critical to prevent leakage of sensitive data from synthetic datasets and to ensure that the original entries cannot be regenerated from the generated data with partial information.

Table 5 evaluates the performance of synthetic data generated by both models using column-wise density estimation, pairwise column correlation estimation, and MLE estimation (Table 7). However, these results may not be sufficient to evaluate synthetic data's overall density estimation performance. The generative model might only learn to estimate the density of each column individually rather than the joint probability of all columns. Additionally, MLE tasks may overlook unimportant columns, leaving to an incomplete reflection of the overall density estimation performance. Therefore, metrics focusing on the entire data distribution, such as the joint distribution of all columns, are adopted.

In Table 8, we present the scores of α-precision and β-Recall [37]. α-precision assesses the fidelity of synthetic data, indicating whether each synthetic example resembles the real data distribution. Fidelity is crucial for measuring how closely the generated samples resemble real ones. High-

Table 7 AUC and RMSE score of MLE by using XGBClassifier and XGBRegressor resp. indicates that the higher the score, the better the performance for AUC and the lower the value is better for RMSE. TabSyn performed better for almost all the dataset

Dataset/Models	Diabetes (AUC)	Obesity (AUC)	Adult (AUC)	Insurance (RMSE)	Asthma (AUC)	Heart (AUC)
TabDDPM	0.677	**0.998**	0.680	1.139	Couldn't generate different class of an attribute	0.573
TabSyn	**0.827**	0.996	**0.908**	**0.434**	**1**	**0.988**

Bold values indicate the better output between the two models being compared

Table 8 Comparing the α-**precision** and β- **recall** scores for synthetic data generated by TabDDPM and TabSyn reveals that TabSyn excels in alpha-precision and beta-recall

Dataset/Models		Diabetes	Obesity	Adult	Insurance	Asthma	Heart
TabDDPM	α-precision	0.655	0.897	0.431	0.55	0.552	0.53
TabSyn		**0.978**	**0.975**	**0.987**	**0.97**	**0.901**	**0.84**
TabDDPM	β - Recall	0.0003	**0.380**	0.008	0.0007	0.0000	0.001
TabSyn		**0.566**	0.304	**0.47**	**0.54**	**0.62**	**0.22**

Bold values indicate the better output between the two models being compared

fidelity synthetic datasets should contain realistic samples. β-Recall evaluates the coverage of synthetic data, indicating whether it can cover the entire distribution of the real data. β-recall assesses the diversity of the generated data, determining whether the samples are diverse enough to cover the variability of the real data.

TabSyn achieves the highest α-precision scores on all the datasets, demonstrating its superior ability to generate synthetic data close to real ones. Moreover, in Table 8, TabSyn shows significantly higher β-recall scores for the diabetes dataset than TabDDPM. However, for all other datasets TabSyn has high β value. This indicates that TabSyn-generated data can cover the diverse variability of the real data.

We believe that assessing the generation quality involves two primary considerations: authenticity (α-precision) and coverage of all modes of the real dataset (β-recall). According to this criterion, TabSyn generates data of the highest quality. It not only exhibits the highest fidelity score but also consistently demonstrates remarkably high coverage across all datasets [38].

A basic method is to look for identical matches, meaning records that appear in both the training set and synthetic sets. However, finding identical matches does not necessarily indicate privacy leakage. Just like any dataset can have duplicate records, a synthetic dataset may also have similar occurrences. Furthermore, removing identical matches from the synthetic data can actually compromise privacy, as it reveals which records are present in the training data by their absence in a sufficiently large synthetic dataset. The concept of identical matches is often extended to measuring the distance to the closest records (DCR), the individual-level distances between synthetic records, and their nearest neighbors in the training dataset. A DCR of 0 indicates an identical match. However, this metric alone does not indicate privacy leakage but rather reflects the data distribution. To interpret DCRs meaningfully, they should be compared to their expected values estimated from a holdout dataset. Thus, we calculate the DCR of each synthetic record relative to both the training dataset and an equally sized holdout dataset. The proportion of records closer to the training data serves as a privacy risk measure, with ties split equally between the datasets. If this proportion is around 50%, it suggests the training and holdout data are interchangeable with respect to the synthetic data, supporting plausible deniability. Even if a synthetic record closely resembles a real-world subject, such resemblances can also occur with unseen subjects [39, 40].

For a training and testing set ratio of a:b, the optimal DCR score for synthetic data with a size similar to the training dataset is a/(a+b). For a 90:10 split, the optimal DCR score should ideally be around 0.90. This means that 90% of the synthetic data points should be closer to the training data than to the testing data [12]. This indicates that the synthetic data have a balanced similarity to the training data, maintaining the expected distribution given the 90:10 split. It suggests that the synthetic data are representative of the real data without overfitting, preserving both privacy and fidelity.

DCR Score Lower than 0.90: A score significantly lower than 0.90 suggests that the synthetic data are not capturing the distribution of the training data effectively, indicating potential issues with the quality of the synthetic data.

DCR Score Higher than 0.90: A score significantly higher than 0.90 suggests that the synthetic data are too similar to the training data, which might indicate over fitting and a higher risk of privacy leakage [40].

Table 9 presents the results of the distance to closest (DCR) score. The DCR metric calculates the average Manhattan distance from each synthetic data point to the closest real data point, assessing how closely the synthetic data resemble the real data [12].

The DCR scores depict that for all the datasets, TabSyn is outperforming TabDDPM.

Table 9 Comparing the DCR score for synthetic data generated by TabDDPM and TabSyn shows that the score is closer to 0.5 for TabSyn in both datasets

Dataset/Models	Diabetes	Obesity	Adult	Insurance	Asthma	Heart
TabDDPM	**0.89**	0.93	0.43	0.54	1	**0.43**
TabSyn	0.87	**0.91**	**0.68**	**0.91**	**0.91**	0.38

Bold values indicate the better output between the two models being compared

7 Conclusion and future works

This study embarked on a detailed comparative analysis of two innovative diffusion generative models, TabDDPM and TabSyn, within the context of synthetic tabular data generation. The findings demonstrate that while both models are robust in their capacity to generate high-quality data, TabSyn consistently outperformed TabDDPM across all tested datasets, showcasing superior effectiveness in maintaining statistical similarity, utility, and privacy .

Our research revealed that the adaptability of TabSyn to various data complexities and its proficiency in handling both continuous and categorical data types make it a preferable choice for synthetic tabular data generation in healthcare settings. This is particularly important given the stringent requirements for privacy and data fidelity in this domain.

The superior performance of TabSyn could be attributed to its integrated approach using a variational autoencoder(VAE) combined with a score-based diffusion model. This configuration allows for more nuanced data generation, particularly in capturing the underlying distributions and inter-variable relationships within datasets.

Despite these strengths, we acknowledge limitations in our evaluation framework:

1. Scalability concerns: The scalability of the models to larger and more complex datasets remains an area for further investigation.
2. Healthcare-specific evaluations: Our framework lacks a detailed assessment tailored specifically for healthcare applications, which could provide deeper insights into data utility and privacy in this sensitive domain.
3. Optimization needs: TabSyn's extended training times on larger datasets highlight a need for optimization to enhance efficiency, particularly for extensive data collections.

This study also highlights several opportunities for further research, particularly in the scalability of the models to larger datasets than those currently tested. As data grow in both size and complexity, understanding how these models scale will be essential for future applications.

Another potential direction for future research involves conducting detailed evaluations of these models using real-world healthcare datasets. The development of a specialized evaluation framework specifically tailored for healthcare data could significantly enhance data privacy and utility in this critical sector.

Further development of the diffusion models is also crucial. Enhancements that enable faster computation, better management of sparse data, or stronger privacy protections could significantly enhance their utility for data scientists. In our findings, TabSyn required more time to train on larger datasets than TabDDPM, indicating a need for optimization to reduce training duration for extensive data collections.

In summary, TabSyn's proven capability to generate high-quality, privacy-conscious synthetic data is highly promising for future data science in healthcare and other fields. Continuous improvements in this technology are expected to lead to more innovative applications of synthetic data, improving our ability to derive valuable insights from complex datasets while ensuring the confidentiality of underlying information.

Acknowledgements I extend my heartfelt thanks to Prof. Dr. Enayat Rajabi for his unwavering support, insightful guidance, and patience. I appreciate his contribution in reviewing and guiding the corrections of this paper. Thanks to Cape Breton University Grant for financial support. Lastly, I am grateful to my family and friends for their moral support and understanding.

Author Contributions Enayat Rajabi and Neetu Kumari both contributed to the study's conception and design. Neetu Kumari was responsible for data acquisition, analysis, and writing, while Enayat Rajabi provided comments for the paper's revision. Both authors have read and approved the final version of the manuscript.

Funding Cape Breton University, grant number RISE-81437, has funded this study.

Data availability All data analyzed in this study were sourced from the publicly available UCI Machine Learning Repository and Kaggle.

Code Availability The code is available on GitHub.

Declarations

Conflict of interest The authors declare that they have no potential conflict of interest.

References

1. Wang, Z., Myles, P., Tucker, A.: Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Comput. Intell. **37**(2), 819–851 (2021)
2. Aguirre, J., Yu, J.Y., Yoon, K.H., Cha, W.C.: High similarity and privacy preserving diffusion model approach, Computationally efficient and stable real-world synthetic emergency room ehr data generation (2023)
3. Nikolenko, S.I.: Synthetic data for deep learning, vol. 174. Springer (2021)

4. Yoon, J., Drumright, L.N., Van Der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ads-gan). IEEE J. Biomed. Health Inf. **24**(8), 2378–2388 (2020)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint. arXiv:1312.6114 (2013)
6. Gonzales, A., Guruswamy, G., Smith, S.R.: Synthetic data in health care: a narrative review. PLOS Digital Health **2**(1), e0000082 (2023)
7. Ahmed, N., Schmidt-Thieme, L.: Sparse self-attention guided generative adversarial networks for time-series generation. Int. J. Data Sci. Anal. **16**(4), 421–434 (2023)
8. Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., Karypis, G.: Mixed-type tabular data synthesis with score-based diffusion in latent space. arXiv preprint. arXiv:2310.09656 (2023)
9. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., Rankin, D.: Synthetic data generation for tabular health records: a systematic review. Neurocomputing **493**, 28–45 (2022)
10. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B.: Medgan: medical image translation using gans. Comput. Med. Imaging Gr. **79**, 101684 (2020)
11. Zhang, Z., Yan, C., Mesa, D.A., Sun, J., Malin, B.A.: Ensuring electronic medical record simulation through better training, modeling, and evaluation. J. Am. Med. Inf. Assoc. **27**(1), 99–108 (2020)
12. Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: Ctab-gan: Effective table data synthesizing. In: Asian conference on machine learning, pp 97–112. PMLR (2021)
13. Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: Ctab-gan+: Enhancing tabular data synthesis. arXiv preprint arXiv:2204.00401 (2022)
14. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint. arXiv:2112.10741 (2021)
15. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: a versatile diffusion model for audio synthesis. arXiv preprint. arXiv:2009.09761 (2020)
16. Yuan, H., Zhou, S., Yu, S.: Ehrdiff: exploring realistic ehr synthesis with diffusion models. arXiv preprint. arXiv:2303.05656 (2023)
17. He, H., Zhao, S., Xi, Y., Ho, J.C.: Meddiff: generating electronic health records using accelerated denoising diffusion model. arXiv preprint (2023). arXiv:2302.04355
18. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: Modelling tabular data with diffusion models. In: International conference on machine learning, pp 17564–17579. PMLR (2023)
19. El Emam, K., Mosquera, L., Hoptroff, R.: Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media (2020)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. **33**, 6840–6851 (2020)
21. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Wei, W.: Machine learning for synthetic data generation: a review. arXiv preprint. arXiv:2302.04062 (2023)
22. Dash, S., Yale, A., Guyon, I., Bennett, K.P.: Medical time-series data generation using generative adversarial networks. In: Artificial intelligence in medicine: 18th international conference on artificial intelligence in medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18, pages 382–391. Springer (2020)
23. Lee, D., Yu, H., Jiang, X., Rogith, D., Gudala, M., Tejani, M., Zhang, Q., Xiong, L.: Generating sequential electronic health records using dual adversarial autoencoder. J. Am. Med. Inf. Assoc. **27**(9), 1411–1419 (2020)
24. Rashidian, S., Wang, F., Moffitt, R., Garcia, V., Dutt, A., Chang, W., Pandya, V., Hajagos, J., Saltz, M., Saltz, J.: Smooth-gan: towards sharp and smooth synthetic ehr data generation. In: Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18, pages 37–48. Springer (2020)
25. Wang, S., Rudolph, C., Nepal, S., Grobler, M., Chen, S.: Partgan: privacy-preserving time-series sharing. In: Artificial neural networks and machine learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29, pages 578–593. Springer (2020)
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint. arXiv:2011.13456 (2020)
27. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning, pp 8162–8171. PMLR (2021)
28. Jia, F., Zhu, H., Jia, F., Ren, X., Chen, S., Tan, H., Chan, W.K.V.: A tabular data generation framework guided by downstream tasks optimization. Sci. Rep. **14**(1), 15267 (2024)
29. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning, pp 214–223. PMLR (2017)
30. Paulin, G., Ivasic-Kos, M.: Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. Artif. Intell. Rev. **56**(9), 9221–9265 (2023)
31. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and gans. Mathematics **10**(15), 2733 (2022)
32. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Adv. Neural Inf. Process. Syst. 32 (2019)
33. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. **34**, 8780–8794 (2021)
34. Shaikhina, T., Khovanova, N.A.: Handling limited datasets with neural networks in medical applications: a small-data approach. Artif. intell. Med. **75**, 51–63 (2017)
35. Chahal, H., Toner, H., Rahkovsky, I.: Small data's big ai potential. Center for Security and Emerging Technology (2021)
36. Plesovskaya, E., Ivanov, S.: An empirical analysis of kde-based generative models on small datasets. Proc. Comput. Sci. **193**, 442–452 (2021)
37. Alaa, A., Van Breugel, B., Saveliev, E.S., van der Schaar, M.: How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In: International conference on machine learning, pp 290–306. PMLR (2022)
38. Ling, X., Menzies, T., Hazard, C., Shu, J., Beel, J.: Trading off scalability, privacy, and performance in data synthesis. IEEE Access **12**, 26642–26654 (2024)
39. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. arXiv preprint. arXiv:1806.03384 (2018)
40. Platzer, M., Reutterer, T.: Holdout-based empirical assessment of mixed-type synthetic data. Front. big Data **4**, 679939 (2021)