Survey on Tabular Data Privacy and Synthetic Data Generation in Industry 4.0

Hadi Koubeissy[1,2*], Amir Amine[1], Marc Kamradt[1], Abdallah Makhoul[2]

[1*]BMW Group, Munich, Germany.
[2]FEMTO-ST Institute, CNRS, University of Franche-Comté, Montbéliard, France.

*Corresponding author(s). E-mail(s): hadi.koubeissy@bmw.de; hadi.koubeissy@univ-fcomte.fr;
Contributing authors: marc.kamradt@bmw.de; amir.amine@bmw.de; abdallah.makhoul@univ-fcomte.fr;

Abstract

Synthetic data is an emerging field that solves the raised need for privacy-preserving data sharing and the lack of real data. Tabular data is the most common data type and is widely used in all domains to train machine learning models, especially in the industrial domain for better decision-making and edge case handling, two key points in Industry 4.0. In this paper, we present and evaluate state-of-the-art models for tabular data generation under a proposed taxonomy consisting of statistical models, generative adversarial networks (GANs) based models, denoising diffusion probabilistic models (DDPMs), and large language models (LLMs). Additionally, we propose a revised evaluation taxonomy consisting of three dimensions, including realism, representativeness, and privacy, where we showcase a set of metrics under each category. The results proved that analyzing models based on multiple metrics from each category could ensure a better understanding of the dataset when used for downstream tasks. Finally, we found that models based on GANs are still a solid option in multiple cases. In contrast, models based on LLMs and DDPMs are more promising, and more research should be invested to overcome limitations such as numerical data representation and long training time for LLMs. Our survey serves as a study for existing models and newer directions in the field with guidelines for evaluation that can be applied not only in the industrial domain but in all other domains.

Keywords: Synthetic data, Tabular data generation, Data privacy, Industry 4.0, Evaluation of synthetic tabular data

1 Introduction

The industrial domain is one of the main engines for modern innovation and development of advanced technologies. The adoption of Industry 4.0 standards [1], including the digitalization factor and usage of advanced technologies such as the Internet of Things (IoT), artificial intelligence (AI), big data, digital twins, intelligent factories [2], and cloud computing is transforming the shape of current industries. The next generation of Industry 5.0, with emphasis on human-machine interaction and sustainability through green, digital, and lean manufacturing processes, dictates an increased reliance on data [3, 4]. As it becomes increasingly apparent that "Data is the new oil," the reliance on data-driven decisions will become the core of the industrial use cases, from decision-making and predictive maintenance [5] to supply chain optimization [6]. Good-quality data to train decision models is becoming essential. However, a significant challenge is the presence and quality of data that need to be used in downstream tasks like machine learning, which will become the sole decision-maker in many systems. Data can take multiple forms, such as images, text, point clouds, and tabular data. In the industrial domain, a vast amount of data is generated from different systems and sensors in the form of tabular data. The industrial domain is not the only one that generates and uses tabular data; the most common form of data in real-world applications is tabular [7], such as the data stored in ACID complement databases. Tabular data can refer to any structured data organized in rows and columns, where rows represent an individual record and columns represent keys or attributes. In the abundance of real tabular data, synthetically generated data may only be a trend if we consider that many obstacles are faced when using real data. Some obstacles are due to privacy concerns, strict data-sharing laws, or limited data availability.

Synthetic tabular data generation is becoming essential across all domains, more pronounced in the industrial sector, where machine learning models rely on large datasets for better decision-making and edge case handling. In most cases, existing data may have missing values or insufficient data points to ensure convergence of the model [8]. Synthetic data, on the other hand, provides a solution that enables the creation of high-quality data that can be used to replace or augment existing data [9]. Additionally, with some models allowing conditional generation, the generated data can be further fine-tuned for specific instances where real data can suffer from imbalanced classes [10]. In digital twins, synthetic tabular data may be used to facilitate testing scenarios like rare events in a safe and controlled environment; it can also be used as metadata for assets to track aging property, maintenance, and specific properties.

The challenges facing generation models are ensuring data representativeness, realism, and privacy, with traditional models not able to cover all the variability in the original dataset or pose a risk to the privacy of sensitive fields. Data privacy is critical nowadays, and strict privacy regulations such as the General Data Protection Regulation, the AI Act by the European Commission, and other laws exist. These regulations set strict rules on processing data and impose anonymization of sensitive fields and personal identifiers. Generative Adversarial networks (GANs), statistical models, diffusion models, and transformer-based models have shown impressive capabilities in generating different forms of data. Some of these technologies were already assessed

for usage as image generators in the industrial domain [11], but their usage for tabular data generation is yet to be validated.

The main focus of this paper is to survey existing models under different architectures, focusing on more recent models with newer architectures to check whether they can be adopted in such contexts. Investigating the usability of the mentioned models should be based on measurable metrics that cover multiple dimensions of how we define usability. This includes checking privacy, realism, and representativeness. All mentioned previously should be under a taxonomy-based analysis, where models from different branches can be evaluated on pre-defined dimensions.

After discussing the problems and objectives, we summarize our contribution to: firstly, surveying existing work in data generation techniques under a proposed taxonomy for statistical models, GANs, diffusion, and Large Language Models (LLM) approaches. Secondly, we propose revised evaluation dimensions where we build on top of existing work to group individual metrics into dimensions that can be better interpreted by data scientists tasked with evaluating a synthetic dataset or model. All the experiments will be based on public datasets widely used in the domain to ensure reproducibility and adoption of our work. To our knowledge, no study surveys the methods used for tabular data generation under the industrial domain, focusing on emerging models like diffusion and LLMs evaluated under the proposed taxonomy. Prior works cover some aspects of this paper, and we are building upon their results to provide a systematically based model and evaluation taxonomy.

For the following sections, the paper will be organized as follows:

- Related work in Section 2 will introduce existing surveys in the domain, introduce the proposed taxonomy, and explore existing models inside each category.
- Datasets in Section 3 will present public datasets that can be used for the development and validation of generative models.
- Evaluation metrics in Section 4 will present evaluation approaches and the proposed taxonomy.
- Experiments in Section 5 will present the experiment results and interpretation in Section 6.
- Conclusion in Section 7 will summarize all the results and discuss future work.

2 Related work

2.1 Existing survey in tabular data generation domain

Many surveys were published tackling the domain of tabular data generation; we will present some of them and discuss the main contributions and focuses of each one. The paper entitled "Survey on Synthetic Data Generation, Evaluation Methods" [12] published in 2022, serves as a perfect start for people exploring the field of synthetic data and, more specifically, GANs for tabular data generation. It offers a taxonomy for generating data like Synthetic Minority Oversampling (SMOTE), deep learning-based approaches, and GANs while covering some evaluation aspects with a focus on global evaluation metrics. However, the coverage remains broad and general, and a deeper understanding of technical insights is needed for each method. Additionally, the survey did not focus on evaluation metrics that can be used to understand different aspects of

generated data. Nevertheless, the survey is a valuable resource for researchers exploring the field.

"Tabular and latent space synthetic data generation: a literature review" published in 2023 [13] is another survey in the same domain, where the authors focus on tabular data and latent space methods. The influence of the paper is the taxonomy-based approach, where the authors proposed a unified taxonomy for classifying synthetic data generation across different data types and ML tasks. Concerning the evaluation of such models, the paper presents some methods for assessing data quality while focusing on metrics that are not context-specific. However, it also lacks a deep exploration of technical applications for each method and does not present a standardized evaluation framework.

"Deep Neural Networks and Tabular Data: A Survey" is a recent survey published in 2024 [14], that provides an overview of deep neural networks (DNNs) for tabular data. The authors categorize the approaches into three main groups: data transformation methods that contain methods for transforming data into suitable formats for usage by neural networks, specialized architecture containing hybrid models specifically designed to capture complex features in tabular data, and regularization models that aim to mitigate the overfitting and non-convergence issues with complex data. One of the critical contributions of this paper is the comparison setup between DNN models and classical machine learning methods in real-world datasets; it still needs more coverage of newer architectures for tabular data generation. The research also did not explore evaluation methods outside of machine learning utility.

Other surveys like [15] explore the usage of GANs in intrusion detection systems (IDS) dataset, particularly the NSL-KDD dataset, for the domain of cybersecurity, where the authors evaluate the performance of GAN models like CTGAN, CopulaGAN, and TableGAN across different metrics like statistical and machine learning-based evaluation. In the medical domain, the focus of survey papers like [16] is to evaluate the approaches to generate synthetic health records in a privacy-preserving context, where the explored models under different categories, including statistical and GAN models with a focus on evaluating privacy metrics. Another survey from the medical domain [17] explores the open-source tools that can used in multiple areas of the medical field for data generation, including imaging and tabular. The paper also reviews the methods under the type of each model. Moreover, the paper also mentions the libraries and frameworks available to provide utilities for data generation. In [18], the authors try to address the challenges of interpretability of machine learning models while trying to check Explainable Artificial Intelligence (XAI) techniques [19] in the field of tabular data that can be used for each model type. The authors show that existing tools for XAI that were initially designed for image or text data struggle when applied to tabular data, and there is a need for more improvement in this area. The previous work specializes in certain domains while trying to answer a more specific research problem.

LLMs are increasingly emerging and becoming more integrated outside the prompt-chat context. The exploration of LLMs in [20] studies their application in prediction and data generation areas where LLMs can show better understanding and handling for missing values while preserving the semantic relationship between variables when

compared to traditional models like XGBoost. Besides their excellent text generation capability, LLMs can also understand table structure better, as mentioned by the authors, with models like TABERT [21] and TAPAS [22] that adapt LLMs for table-related tasks. Lastly, the paper also covers the challenges with data preparation needed to transform the tabular data into a suitable format for LLMs.

The existing work offers valuable insights into tabular data generation, focusing on introducing the field and explaining existing methods. However, existing work needs a more specialized focus on industrial applications and their unique challenges, such as handling mixed types, preserving feature interactions, and privacy concerns in the generated data. Our work distinguishes itself by addressing these gaps and proposing a taxonomy that categorizes synthetic data generation methods. Additionally, we survey emerging models in fields like diffusion and LLMs and their potential use in future applications. In addition, we shed light on many evaluation metrics that can provide more explainability of the generated data, an essential aspect studied in industrial domains to ensure predictable and consistent decision-making.

2.1.1 Proposed taxonomy

This section presents a taxonomy for synthetic data generation based on four main groups: statistical models, GAN models, diffusion models, and transformer-based LLMs. The suggested taxonomy allows us to assess models' properties under each category systematically. This structure is based on the architectural differences in each approach. Statistical models rely on probabilistic and statistical models to infer the underlying distribution of the original dataset and try to sample new data points following the same distribution; this makes them efficient and interpretable but less flexible in capturing complex distribution and preserving semantic relationships between variables. On the other hand, GANs use the adversarial learning problem to generate representative data but can suffer from training and stability problems. Diffusion models present a newer approach in this field after proving themselves in image generation tasks by gradually improving synthesis after each iteration. Transfomer-based LLMs are important when preserving the semantic relationship between variables and taking advantage of their embedded knowledge.

2.2 Statistical models

In the following section, we will introduce statistical models, which rely on statistical methods to generate data with the same probability distribution as actual data. The following class of models ensures the explainability of the generated data. In some areas, Bayesian networks can construct a probabilistic graphical model with conditional dependencies between variables [23].

2.2.1 Coupla Flow

Coupla Flow, a statistical model proposed by Sanket Kamthe et al. in 2021 [24], addressed the challenge of probabilistically modeling synthetic tabular data in a robust and interpretable way. While achieving good results, previous GAN-based and Variational Auto Encoders (VAEs) models have difficulty interpreting their generation

results. In addition, the loss function of a GAN model needs to be case-specific to achieve the best results. Moreover, GAN-based models can suffer when modeling mixed-type variables. For the following reasons, Coupla-Flow can be used in scenarios where the model behavior needs to be predictable, like medical and financial domains governed by strict laws. The proposed model combines copula theory with normalizing flows to create the synthetic data generator. The modeling flow learns the marginal distribution and transforms it into a uniform distribution. Afterward, they estimate the Copula density, sample from the uniform distribution, and apply the learned Copula for transformation into correlated uniform marginals. Finally, the model transforms the sampled data into the original data space. Applying this modeling ensures that the generated data retains the statistical properties and dependencies of the real datasets. The authors evaluated Copula Flow in several aspects: mixed-variable learning, density estimation, and machine learning utility. They trained Gaussian Mixture (GM) and Bayesian Network (BN) models on the synthetic data to validate the model's utility and measured the log-likelihood. While Copula Flow produced results close to those of GAN and TVAE models, it did not significantly outperform them. However, Copula Flow is ideal for applications where privacy and model interpretability are critical.

2.3 GAN based models

Generative Adversarial Networks (GANs) introduced by Ian J. Goodfellow et al. in 2014 [25] are a class of generative models influential in the generative AI field. The basic architecture comprises two neural networks: A generative model G and a discriminative model D. Generator G captures the data distribution and generates new data samples. At the same time, Discriminator D evaluates the probability of the generated sample coming from G or from the real data in a minimax two-player game. During the training phase, Generator G starts with random noise as input and tries to generate data that are as realistic as possible. At the same time, Discriminator D takes both the real example and the generated example and tries to distinguish between them. The final objective of the model is for Generator G to generate data that cannot be distinguished by Discriminator D. Multiple improvements to the original GAN architecture were introduced to solve and adapt the model to different use cases, one of which is called Wasserstein GANs, proposed by Martin Arjovsky in 2017 [26]. WGAN tries to solve the mode collapse and training instability associated with vanilla GAN. The key feature of WGAN is the usage of Wasserstein distance as a measure of the similarity between real and generated data, which leads to smoother gradients and stabilizes the training.

Conditional GANs (CGANs) [27] is an additional extension of GANs that enable conditional generation. They allow the generation of new instances that fall under specific class labels or data attributes. A conditional vector is introduced to achieve conditional generation, which contains additional information for the Generator G and Discriminator D to influence the generation process. In addition to generating conditional data that falls under a particular distribution and controlling the data diversity within the specified conditions, CGAN can also be used to handle mixed data types found in some tabular data by conditioning the generation process on the

relevant attributed to ensuring the exact structure of the generated data as the original data.

Deep Convolutional GANs (DCGANs) introduced by Alec Radford et al. [28] in 2016 incorporate convolutional neural layers and deconvolutional layers, replacing the fully connected layers, which improved the quality of the generated data by better capturing the spatial hierarchies of the data, which is beneficial for image application and was later adopted in tabular data generation.

In the next section, we will present some influential models based on GAN architecture.

2.3.1 TGAN

TGAN proposed in the following paper [29] is a tabular generation model based on GAN architecture that addressed the challenges of generating mixed-type data with a distinction on multinomial variables that needed to be adequately tackled in previous scientific work. Multinomial variables correspond to multi-value categorical data in contrast to binary values. Additionally, the model employs Long Term Short Term Memory (LSTM) with an attention mechanism to better capture inter-variable correlation and handle the sequential generation of columns in a table. The other improvements in the model architecture can be summarized by employing reversible data transformation, like mode-specific normalization, such as Gaussian kernel density estimation, to estimate the number of modes in continuous variables with the help of the Gaussian Mixture model. The categorical data type is transformed using one-hot encoding for multinomial variables, adding noise to binary variables, and then using Softmax to generate the probability distribution. The authors also compare TGAN to Table-GAN and distinguish both implementations by their neural network architecture and objectives. While TGAN uses LSTM with attention to generate data column-by-column, aiming to learn the marginal distribution of each column by minimizing Kullback-Leibler (KL) divergence in order to improve the handling of complex dependencies between mixed data types. In contrast, TableGAN uses a Convolutional Neural Network (CNN) to generate complete data records at a time. Moreover, the optimization of prediction accuracy is done by minimizing the cross-entropy loss. The differences presented are tailored toward preserving the correlation between original tabular data, as shown in the evaluation section of this paper. The evaluation used falls under two metrics:

1. Machine learning efficacy is measured by training different models on the real and synthetic datasets and then measuring the accuracy between the two models.
2. Normalized mutual information (NMI) is used to evaluate the correlation between real data and how much the model preserves this correlation in the synthetically generated data.

In both evaluations, TGAN outperformed Gaussian Copula and Bayesian networks in both evaluation dimensions.

2.3.2 CTGAN

The authors of this paper in [30] propose a new model that targets the modeling of joint probability distributions of all columns in datasets that contain both discrete and

continuous values. The referenced Bayesian and older GAN models [29] approaches struggle with mixed-type data, especially the challenge of imbalanced discrete columns. The proposed approach consists of introducing a conditional generator and training-by-sampling. The conditional vector is implemented to overcome the class imbalance problem found in discrete columns, one-hot encoding, and a mask vector representing the discrete values in the datasets to express conditions more effectively and avoid the class imbalance problem in some datasets.

Introducing a conditional vector allows the generation of synthetic data that adhere to specific conditions set for the generation process. To assess the output produced by the conditional generator, training by sampling is introduced. This method samples the real training data and constructs the conditional vector (cond), which helps the critic estimate the distance between the generated and real data distribution.

To better represent the dataset in the latent space, a mode-specific normalization is used to represent the distribution of continuous values that follow more complex distributions that are not necessarily Gaussian distributions. Other approaches used min-max normalization that might fall shorthanded in representing more intricate distribution found in continuous values.

The evaluation was done on the simulated dataset and six commonly used machine learning datasets from the UCI machine learning repository [31]. The two metrics used to evaluate the generation quality of CTGAN are likelihood fitness and machine learning accuracy. Likelihood fitness was used to evaluate if a synthetic dataset follows the same distribution as the simulated datasets. The application of this metric was only possible on the simulated dataset as the distribution is already known, while the distribution of real-life datasets is unknown. On the other hand, machine learning efficacy was used on the UCI datasets by training various machine learning models on both real and synthetic datasets and then evaluating the models' performance using machine learning metrics like accuracy and precision.

2.3.3 Table-GAN

Table-GAN proposed in [32] is a model designed to address several critical issues in synthetic data generation, particularly concerning privacy and maintaining the original data distribution. The primary focus of Table-GAN is to ensure the privacy of data sharing. It effectively counters various privacy attacks, such as re-identification attacks, attribute disclosure, and membership attacks. This is achieved through its inherent GAN design that produces entirely new records, unlike the training data, thereby minimizing the risk of privacy breaches. To further enhance privacy, Table-GAN employs hinge loss during the generator training. Hinge loss introduces a margin in the decision process, which slightly disrupts the training and helps protect against privacy attacks. This technique is widely used in Support Vector Machines (SVMs) to create a decision boundary margin.

The model supports categorical, discrete, and continuous data. Future work aims to extend support to other data types, such as strings, making the model more versatile and applicable to a broader range of datasets. Table-GAN consists of three main components: a Generator, a Discriminator, and a Classifier Network. The classifier network is crucial for maintaining the semantic integrity of the generated records.

The loss functions used in Table-GAN include the original DCGAN loss, chosen due to the maturity and reliability of the DCGAN model in the state-of-the-art (SOTA) at the time of publication. Additionally, information loss measures the discrepancy in statistical properties (mean, standard deviation) between the real and synthetic data to preserve the original data distribution, and classification loss is used by the classification network to ensure that the semantic integrity of the generated data is maintained.

Table-GAN is evaluated based on two main criteria: privacy, assessed by the distance to the closest record (DCR), ensuring the generated data does not closely resemble any single real data point, and data utility, evaluated through statistical comparison and machine learning score similarity. This includes using various classifiers such as Decision Trees (DT), Random Forests (RF), Adaboost, Multilayer Perceptrons (MLP), Linear Regression, Lasso, Passive/Aggressive Regressor, and Huber Regressor to compare the performance of models trained on synthetic data versus actual data. Additionally, Table-GAN provides parameters that allow users to balance the trade-off between privacy levels and model comparability. This flexibility ensures that users can tailor the model to meet specific needs and constraints in various applications.

In conclusion, Table-GAN offers a robust solution for generating synthetic data while preserving privacy and maintaining data integrity. Its sophisticated architecture and comprehensive evaluation methods make it a valuable tool for researchers and practitioners in synthetic data generation.

2.3.4 CTAB-GAN

The authors of CTAB-GAN [33] as previous models like [29] also target synthetic data generation with more focus on data privacy while implementing a conditional GAN-based model that's advantageous when generating synthetic records following specific conditions. In terms of contribution and novelty compared to previous models, CTAB-GAN added support for mixed data type variables (meaning columns containing both categorical and continuous values or continuous and missing values) by introducing a mixed-type encoder that encodes mixed-type variables as value-mode pairs. VGM (Variational Gaussian Mixture) estimates the total number of modes and then fits a Gaussian mixture to represent the continuous part of the variable. Continuous variables of the distribution are normalized with the mode having the highest probability, and finally, categorical values are represented using a one-hot vector. Missing values are also represented as a unique class, enhancing the encoding process of the mixed-type variables.

The model also implements information loss as a penalization function for the generator to preserve the same statistical distribution between the generated and real data. Moreover, classification loss is introduced as another penalization function to preserve the semantic relation between variables within the same record for the generated data, a common problem with GAN-based models. The long tail issue can be expected in real-life datasets, and it refers to data points that can be distant from the more dense data region in the dataset. VGM can struggle to encode those data points, so the authors implemented pre-processing by implementing a logarithm transformation that reduces the distance between those points and the majority of the

distribution. Conditional vector and training by sampling are also used in training to overcome imbalanced training datasets and allow for minority classes to be presented in the conditional vector, thus in generated data. To create the conditional vector, the model chooses randomly but with an equal chance variable and then calculates the probability distribution of each mode or class within the selected variable. Afterward, a logarithm transformation is applied to give a higher weight to the minority class or mode. The usage of this algorithm worked as a mitigation to the collapse issue found in GAN-based models because the model does not focus on common data points while ignoring minority and distant points.

In terms of evaluation of the model, they used three UCI datasets - Adult, Covertype, and Intrusion- along with Credit and Loan, which totaled to five datasets. They used CTGAN, TableGAN, CWGAN, and MedGAN for the baseline comparison. The evaluation metrics used to assess the quality of the generated data are (1) machine learning utility, (2) statistical similarity with metrics like Jensen-Shannon Divergence (JSD), Wasserstein distance (WD), and Difference in pain-wise correlation (Diff. Corr) were used, (3) privacy preservability metric to measure if the generated data preserve privacy, DCR and Nearest Neighbour Distance Ratio (NNDR) were used; Both metric tries to find the distance between a given generated point and nearest neighboring real point(s). The evaluation showed CTAB-GAN outperforming the mentioned models in terms of ML utility, statistical similarity, and reasonable in the privacy-preserving area, especially when compared to Table-GAN.

2.3.5 CTAB-GAN+

CTAB-GAN+ comes as a successor of CTAB-GAN [34] with more focus on privacy-preserving aspects and improved support for mixed-type variables compared to CTAB-GAN. The authors adjust the generator architecture by adding a downstream loss in addition to generator loss and information loss. Downstream loss is implemented to link independent and target variables to preserve the semantic correlation between variables in each record. The model's training process included Wasserstein loss with gradient penalty to ensure that the discriminator generates more bounded gradient norms. This stabilizes the training by avoiding gradient exploding or vanishing. To implement this, a gradient penalty was added to the Wasserstein loss, forcing the gradient to stay bounded around one. Clipping the gradient near one prevents the discrimination from becoming too weak or dominant, stabilizing the training and improving the convergence. Another improvement over CTAB-GAN, specifically in the mixed-type encoder, is the integration of mode-specific normalization (MSN) alongside VGM for the continuous part of mixed-type variables; CTGAN inspired this idea [30]; MSN offers better normalization for each distribution, yielding better-encoding precision. Additionally, differential privacy stochastic gradient descent (DP-SGD) was implemented to add noise to the gradient in the training phase to avoid any data leak into the generated data by eliminating the influence of a single data on the outcome but giving more influence to the whole data. The evaluation results were based on two protocols with and w/o DP. For w/o DP protocol, the goal was to generate data that was as realistic as possible without privacy constraints, and evaluation was based on machine learning utility and statistical similarity. Specifically, CTAB-GAN +

improved AUC by 56.4% and accuracy by 41.2%. Under differential privacy, the model outperformed models within this category, such as PATE-GAN and DP-WGAN, in both machine learning utility and statistical similarity.

2.3.6 Casual TGAN

Generating tabular data introduces a challenge that many models try to tackle. One specific challenge is modeling the inter-variable relation, including causal dependencies, in the newly generated data. Causal-TGAN (Causal Tabular Generative Adversarial Network) proposed by [35] addresses these challenges by incorporating causality into the generative process.

Domain knowledge is incorporated to construct the causal graph; this can be fully or partially known, allowing the model to incorporate varying expert knowledge. Moreover, to overcome the generation problem in scenarios with partial causal knowledge, the model uses a causally-aware generator for known relationships and a conditional GAN that's used in [30] for the remaining variables. If the causal relations are wholly known for the training process, structural causal models (SCM) are used to model the relationships between variables using a multilayer perceptron (MLP). In contrast, in partial mode, the Causal TGAN first fits on a subset of known causal relations, and then conditional GAN is used to generate the other subset. In both cases, WGAN loss with gradient penalty is used for the training process. However, in the case of partial knowledge, the parameters of the causal generators are frozen after fitting, and only conditional GAN parameters are updated. Moreover, mode-specific normalization was also used to normalize continuous, discrete, and mixed-type variables.

The authors used both simulated and real data for evaluation. The simulated data built by Bayesian networks have full causal graph knowledge, while real-world datasets have partial knowledge. Kullback-Leibler divergence (KLD) and Log-Cluster (LC) are used for the simulated datasets, and machine learning efficacy (MLe) is used for the real-world datasets. The results showed that Causal-TGAN outperformed other models in comparison to MedGAN, TableGAN, TVAE, and CTGAN in all tests, which shows the importance of causal architecture in generating data that preserves the interrelation between variables.

2.4 Diffusion based models

Diffusion models are generative models used to create new data samples by stimulating a diffusion process. In a diffusion process, the data undergoes a series of transformations, which involve adding noise and then denoising. This process aims to understand the underlying distribution and characteristics of the dataset features, and then running the denoising process will generate new samples from the noise. The noise is incrementally added to the data over each step in the forward process. This process implies a noise distribution where many algorithms, such as Gaussian noise, can be used. In the denoising process, the main aim is to denoise the data incrementally and reconstruct the original data. The denoising process is achieved by learning a series of denoising functions that predict the original data from their noisy counterparts. The training objective ensures that the model captures the data distribution and reverses

the diffusion process during the generation phase. The next section will introduce the modeling of tabular data following the denoising diffusion models, where the adoption of diffusion models might be more relevant, especially after they have proven themselves in the generative imaging area.

2.4.1 Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPM) introduced in [36, 37] are a specific type of diffusion model that focuses on probabilistic denoising. DDPMs transform data into noise through a forward process and learn to reverse this transformation to generate new data. In the forward diffusion process, data is progressively noised over multiple timesteps until it resembles a standard Gaussian distribution. This gradual addition of noise destroys the information in the data, making it increasingly indistinguishable from pure noise. The reverse diffusion process involves denoising the noisy data back to its original form by learning a sequence of denoising autoencoders parameterized by a neural network. The model learns to approximate the original data distribution from the noisy versions by predicting and removing the noise added during the forward process. The training of DDPM involves minimizing the difference between the actual noise added and the noise predicted by the model at each step. This approach results in a stable and effective training process, unlike adversarial training used in GANs, which can be unstable. DDPMs are known for their high-quality and diverse data generation capabilities, making them suitable for various applications such as image synthesis, speech generation, and tabular data generation.

TabDDPM: "Modelling Tabular Data with Diffusion Models" authored by Akim Kotelnikov et al. published in 2022 [38]. This paper falls under the generation and modeling of tabular data. This paper explores the application of DDPM to tabular data to improve the quality of synthetic data generation over existing models like GANs and VAEs. The model is targeted for mixed-type data containing numerical and categorical features. Another aspect of this paper is the adaptation of diffusion models for tabular data after the success of such models in other domains like image generation and speech processing.

The proposed model integrates separate diffusion processes for handling numerical and categorical features. Numerical data are handled by Gaussian diffusion and encoded using Gaussian quantile transformation, where each numerical feature undergoes a noise addition process using Gaussian distribution. The noising process is applied in small increments in each forward step. For this purpose, Gaussian distribution can allow for the incremental noising process without distorting the features in a non-reversible way. A one-hot encoder is used to encode the categorical features. Then, multinomial diffusion is applied to corrupt the data by applying uniform noise to all the categories. As categorical data represent discrete values, multinomial distributions are appropriate because they can represent the probability distribution over a finite number of categories. In addition, the multinomial diffusion process adds uniform noise, which ensures that all the categories have an equal chance of being selected as the noise increases, meaning that no category will be under-represented or over-represented. Gaussian and multinomial diffusions effectively support heterogeneous data types in tabular data. In the denoising process, Multi-Layer Perceptron (MLP)

is used to denoise the data step-by-step, transforming the noisy data into synthetic data by predicting the original data points from the noisy version at each diffusion step. More technically, MLP will produce two outputs corresponding to the type of diffusion process. For the numerical features, MLP predicts the noise component (ϵ) that will be subtracted from the noisy data to get the denoised data, while for the categorical data, the MLP will predict the probability of each category to reconstruct the denoised data. The loss function of TabDDPM involves combining two losses: a loss for numerical features, which follows the standard loss objective of DDPM, meaning the mean squared error (MSE) between the predicted noise and the actual noise over the timesteps, and a loss for the categorical features using KL convergence between the original probability and the model predicted probability. Both numerical and categorical losses are combined and normalized by the number of features found in the dataset.

The evaluation was conducted against other existing generation models like TVAE, CTABGAN, CTABGAN+, and SMOTE while using machine learning efficiency and privacy metrics as evaluation metrics. For the machine learning efficiency, DT, RF, and Logistic Regression (LR) were used to train models based on real and synthetic data and then evaluated by F1-score for classification, and R^2 was used for the regression performance. In addition, another protocol was used for the machine learning efficiency by using specialized models like CatBoost and specialized MLP tuned on each evaluated dataset, specifically using the search spaces approach from [39]. For the privacy aspect of the evaluation, the authors used the DCR metric that measures the proximity of synthetic data to real data points. The datasets used are 15 real-world public datasets with different data types (numerical, categorical, mixed-type). Regarding machine learning efficiency, in some tests, TabDDPM performed better than TVAE, CTABGAN(+), and SMOTE. For privacy, the authors compared TabDDPM to SMOTE only, showing a better privacy-preserving aspect.

2.5 LLM based models

Large language models (LLMs) dominate today and pioneer language understanding by relying on billions of training model parameters. With the introduction of numerous LLM families and Generative Pretrained Transformers (GPTs), especially OpenAI's ChatGPT GPT-4 [40] and GPT-4o. Transformers are the building block of most LLMs, as introduced in [41] the proposal of a self-attention mechanism that enables the calculation for each token in a parallel way an "attention score" and understanding the influence of each word in a sequence to predict the next token in a generation scenario. This architecture allows an unprecedented parallelization on GPUs, which allows billions of trainable parameters to be calculated to understand the language effectively. These large pre-trained models (PLMs) can be later fine-tuned for other downstream tasks while taking advantage of embedded knowledge within those models. As suggested in [42, 43], early encoder-only PLMs such as text classification were used for language understanding. BERT (Bidirectional Encoder Representations from Transformers) [44], and other variants from the same family, like Roberta, enabled embedding word sequences in latent space where each word represents a high-dimensional vector. The advantage of BERT lies in its bidirectional

training, which allows it to understand the context of words more effectively by considering both preceding and succeeding words, thereby improving the model's ability to capture relationships between similar and opposing words. Decoder-only PLMs like GPT-1 [45] and GPT-2 [46] developed by OpenAI focus on the decoder part of the transformer architecture, the introduction of diverse unlabeled text for general language understanding and the ability to fine-tune it for specific tasks in GPT-1. At the same time, GPT-2 expanded this approach on the WebText dataset and included some model enhancements that allowed for an increased context size. These advancements laid the ground for successor models like GPT-3 and GPT-4. The decoder-only models, also called auto-regressive models, predict the next word in a sequence of tokens, which makes them suitable for text-generation tasks. Encoder-decoder models [47] reformulate the problem as a sequence-to-sequence generation task, meaning a unified block that understands and generates the text; BART [48] is a pre-trained model that follows this architecture and is trained by corrupting text using noise function and reconstructs the original text. Because those models' ability to access the entire sequence of words in the encoder and the decoder can transform the generated vector from the encoder into an output sequence, they are suitable for more complex tasks where an input sequence needs to be transformed, like the case of translation. The following introduction about LLMs and their overview and general functioning inspired the adaptation of their pre-trained models and architecture for tabular data generation tasks.

The following section will introduce recent approaches to integrating LLMs adapted for tabular data generation, especially how to overcome some challenges with encoding tabular data in formats suitable for LLM fine-tuning.

2.5.1 GReat

Multiple approaches to adopt LLM models into tabular data generation harnessing the power of language models as natural data generation tools. One of those approaches solutions named **GReaT** (**G**eneration of **Rea**listic **T**abular data) published in 2023 by Vadim Borisov et al. [49]. The authors of this paper try to explore the potential of auto-regressive LLM in generating synthetic tabular data. The proposed model tries to overcome the limitations of existing generative models, such as the extensive preprocessing required for the data, which includes encoding categorical data, normalizing data points, and handling missing values and outliers. Additionally, most generative models do not preserve or take advantage of existing contextual information in the dataset like in the mentioned dataset of Adult Income dataset [50] where the pair of features $age, marital - status, education$ have an underlying direct correlation that most other models do not consider it due to the processing that removes this semantic correlation. Arbitrary conditioning is another focus of the proposed model, which allows the conditioning of a subset feature or attribute without retraining. In simpler terms, arbitrary conditioning allows the user to specify a conditional generation by specifying some initial generation conditions, for example, generating samples with conditions like $age = 45; marital - status = Married - civ - spouse$.

The proposed model starts by training the pre-trained LLM on the textual encoding version of the training data, meaning transforming the tabular information into a

textual version suitable for the LLM tuning as shown in Fig. 1 and addition feature permutation for better generalization of the model, and allows it to generate any subset of data based on conditioned input. It is worth mentioning that the textual encoding is also applied to numerical values, meaning they are represented as character sequences, which LLMs are capable of interpreting [49, 51, 52]. One of the benefits of textual encoding is preserving the feature name and the underlying semantic meaning that LLM can interpret and understand based on the embedded knowledge from its context, which is influential in the generation process compared to previously discussed models that ignore the informational context in the feature names. The sampling process in GReaT can be controlled following either: (i) feature name preconditioning, where only one feature name is specified, for example, *age*, (ii) name-value preconditioning, where a feature and specific value can be specified, example $age = 45$, and (iii) multiple name-value pair preconditioning where multiple features conditions can be specified like $age = 45; marital - status = Married - civ - spouse$.

The evaluation process was based on six datasets from Kaggle [53] and the UCI repository [31], with baseline models TVAE, CopulaGAN, and CTGAN. The trained LLM models used are Distill-GReaT based on a distilled version of GPT-2 named DistilBERT [54] with 82 million parameters, and larger context model GPT-2 [46] with 355 million parameters. The evaluation was based on machine learning efficiency measured based on LR, DT, and RF results. Furthermore, the DCR histogram was used to measure the similarity of the records to the actual dataset. In both aspects of the evaluation, both models, Distill-GReaT and GReaT, overcame reference models. The advantage of the GReaT model is an introduction of LLMs into the tabular synthetic data generation domain and proposing an alternative for other generation models with their associated problems like the problematic of data encoding in high dimensional space, also known as *the curse of dimensionality* mentioned by Richard E. Bellman [55]. Furthermore, LLM models and their underlying auto-regressive architectures can generate semantically related content based on sequential input, maintaining the semantic coherence of the features. Conversely, the training process can be lengthy in terms of time and processing power. For example, a fine-tuning job for GReaT can take up to 9 hours compared to 1 minute in CTGAN, and the sampling process can also be approximately 377 times slower than CTGAN.

2.5.2 Tabula

TabuLa is another proposal from the authors Zhao et al. in 2023 [56] that aims to improve existing tabular data generation LLM models. The proposed model focuses on improving the fine-tuning process and model selection by using a randomly initialized state foundation model. Afterward, the selected model can be trained for tabular generation tasks on a given transformed data, enabling this model version to be fine-tuned for other datasets. The selection of more specialized models and the proposed fine-tuning process aim at creating a faster and more accurate synthesizer that can also be used for various tabular data synthesis tasks. The start point of the foundation models a randomly initialized language model instead of a general language processing model like GPT-2. The idea behind this approach is to utilize the ability of the large models to understand semantics without the usage of the

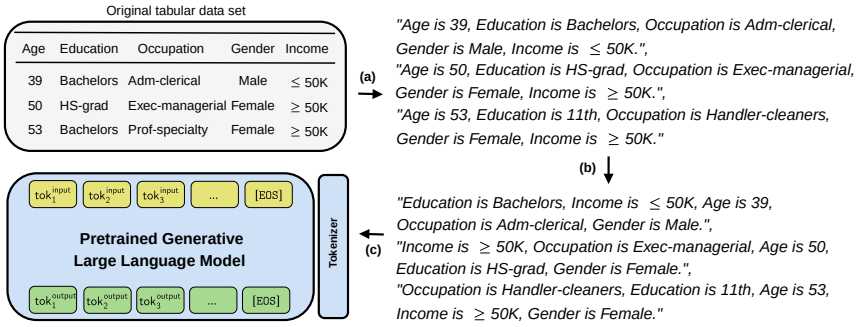

Fig. 1: GReaT proposed textual encoding for LLM fine-tuning. First, the tabular data are transformed into a textual representation suitable for LLMs, and then feature order permutation is applied for better generalization. Data features are from [50], figure taken from [49].

learned NLP capabilities. On the side of processing and tokenization, TabuLa uses more simplified processing than GReaT to transform the training data into trainable examples. While the predecessor model tries to construct a sentence from a row (i.e., $age\,is\,45\;and\;marital-status\,is\,Married-civ-spouse$), TabuLa uses a shorter representation of the same data point example for clarity ($age\,45\;marital\,MCV$). We can also notice that the feature names and values are compressed into one token for increased performance. In cases where conditional sampling is not required, TabuLa does not employ feature permutation and applies middle padding. Padding is a technique applied by tokenizers in LLMs to add non-representative padding to the token sequence of each training sample for optimized batch processing. Left and right padding are standard in tokenization, each having its uses like emphasizing sequence position or aligning all examples to the same length; middle padding is used in TabuLa to preserve the key-value pair position across all the training data.

The evaluation process was done on two aspects: MLP and statistical similarity, TabuLa had better results than reference models in MLP, and the statistical similarity between different TabuLa models: $\boldsymbol{TabuLa_P}$ initialized from pre-trained DistilGPT-2, $\boldsymbol{Tabula_F}$ fine-tuned from $\boldsymbol{TabuLa_P}$ on Intrusion dataset, ($\boldsymbol{TabuLa_P}$, $\boldsymbol{Tabula_F}$ both connotate predecessor approach in training tabular models) and $\boldsymbol{TabuLa_R}$ initialized on random DistilGPT-2, and $\boldsymbol{Tabula}$ fine-tuned from $\boldsymbol{TabuLa_R}$ on Intrusion dataset. Those models' results in terms of loss and statistical similarity of the generated data show that starting from randomly initialized LLM models and fine-tuning based on that model from a tabular dataset always achieves better results and faster convergence than starting from a fully knowledgeable NLP-targeted foundation model.

3 Datasets

Evaluating the performance of a synthetic data generation model is a crucial aspect of our research. It requires a structured approach to ensure that the generated data

meets the desired criteria of accuracy and usability. Given the importance of real-world relevance and reproducibility, it is essential to base this evaluation on carefully selected datasets. Indeed, the evaluation process of the generative models is based on the datasets used, each of which represents a real-world use case or problem to be addressed. To ensure a fair evaluation of the model and facilitate reproducibility across different setups, publicly available datasets are employed. These standardized datasets help ensure consistency in the evaluation and comparison process across various experiments.

In ecology, the Abalone dataset [57] from the UCI repository contains 4,1777 instances that represent the age and physical characteristics of abalone. Similarly, Covertype dataset [58] can be used to classify seven forest cover types with more than 581K instances.

In the healthcare field, it exists many datasets and health surveys such as the Cardiovascular Disease dataset and the Diabetes dataset from Kaggle, that facilitate classification tasks in the predictive health domain and medical diagnostics. Datasets in the financial field range from house pricing prediction, like California Housing, to insurance datasets, like the car insurance dataset from the bn-learn repository [59] based on the Bayesian Network to generate records. A loan dataset is also available to predict eligibility based on customer information.

For customer behavior and recommendation systems, many datasets exist to study and predict the behavior of customers based on multiple features, such as the Churn dataset. Additionally, customer reviews from social media can be studied using datasets like Expedia Hotel dataset [60], and Facebook comment volume [61]. In the domain of cybersecurity, the intrusion dataset [62] for detecting network intrusions.

The presented examples demonstrate the practicality of the research, enabling the benchmarking of tabular generation models across different domains. They can support a wide range of tasks, from classification and regression to recommendation systems. In industrial applications, where there is no standard dataset due to the scarcity of applications and diversity of the domain, these examples can serve as a solid starting point for evaluation. We limited our experiments to publicly available datasets, testing various datasets that are suited for multiple applications and exhibit diverse targets and characteristics. Table 1 summarizes some available and widely used datasets with several instances, target problems, and descriptions.

Table 1: Summary of datasets for tabular data generation evaluation.

Dataset Name	Source	Description	Instances	Problem Type
Abalone	UCI	Predicting the age of abalone from physical measurements.	4,177	Classification/Regression
Buddy	Kaggle	Pet adoption dataset based on physical aspects and breed.	26,906	Classification
California Housing	Kaggle	Housing prices in California based on various factors.	20,640	Regression
Cardiovascular Disease dataset	Kaggle	Medical records for cardiovascular disease detection.	70,000	Classification
Churn Modelling	Kaggle	Customer information to predict if a bank client closed his account or not.	10,000	Classification
Facebook Comments Volume	UCI	Predicting the number of comments a post will receive on Facebook.	40,949	Regression
HIGGS	UCI	Distinguishing between signal processes producing Higgs bosons and background processes.	11 million	Classification
House Pricing	Kaggle	Real estate pricing prediction using various features of residential homes.	2,919	Regression
Intrusion	UCI	Network intrusion detection.	4 million	Classification
King	Kaggle	House sales data from King County, USA.	21,613	Regression
Bank Loan Modeling	Kaggle	Predict loan eligibility based on customer details.	5000	Classification
Covertype	UCI	Predict forest cover type from different attributes.	581,012	Classification
Fake Hotel Guests	Kaggle	Classifying fake hotel guests based on booking details.	2,000	Classification
Insurance (Bnlearn)	Bnlearn	Modeling risk for car insurance policies.	-	Classification/Regression
Insurance (Kaggle)	Kaggle	Predicting health insurance costs.	1,338	Regression
MNIST28	Lecun	Handwritten digit recognition.	-	Classification
Child	Bnlearn	Health diagnosis through a diagnostic Bayesian network.	-	Classification
Adult Income	UCI	Predicting whether income exceeds $50,000/year.	48,842	Classification
NHANES Health	CDC, Kaggle	Tracking health conditions.	-	Classification/Regression
Expedia Hotel Logs	Kaggle	Predicting which hotel cluster a user is likely to book.	-	Multi-case
Online News Popularity	UCI	Predicting the popularity of online news articles.	39,797	Classification/Regression
Home Equity Line of Credit(HELOC)	Kaggle	Predicting credit risk for Home Equity Line of Credit.	10,000	Classification
Diabetes	Kaggle	Diagnosing diabetes in Pima Indian women.	768	Classification

4 Evaluation metrics

Evaluating the performance of synthetic data generators (SDGs) is crucial before using a synthetic model for data generation. This can ensure that the produced data that will be used for later machine learning tasks is considered: (i) reliable, meaning it can be a trustable source for replacing the real data or be used as an augmentation source for the existing real data, (ii) preserve the privacy of the original real data and do not expose any trained on data records or sensitive information, (iii) represents the original data in terms of distribution, and statistical aspect, (iv) diversity and novelty, providing a variety and diverse of synthetic instances without causing overfitting. Many categories were proposed to group individual metrics under categories; those categories can be used to describe certain aspects of the generated data and evaluate the model performance. Evaluating the model against accuracy and traditional machine learning metrics like F1-score or R^2 validates that the usage of the synthetic data can achieve the same level of application performance as the real data when used to train the same ML. The following methodology does not ensure privacy or direct similarity to the real data. In previous work done on the evaluation metrics in this domain [63], the authors proposed a taxonomy based on a broad evaluation including univariate fidelity, which measures the structural similarity between synthetic and real datasets, bivariate fidelity that measures the correlation between variables of the same row, population fidelity in which they try to find the distribution of a masked part of the dataset that was obfuscated to mesure the models' understanding of the dataset without relying on some details in the data like sensitive fields. A narrower application fidelity is also mentioned, where the SDG models are evaluated in the final application domain. Other research, like [64], proposed a similar taxonomy that is based on representativeness, measuring the distribution of the data and novelty, measuring how novel and new the generated records are, in addition to the realism measure that can be based on human evaluation for specific application field, diversity that tries to find if repetition of the same data instance is present, and coherence which might be more relevant in sequential data were some orders needs to be present in the data. In more specific fields like the medical field, a more application-specific approach for identifying metrics and categorizing them was suggested in [16, 65] for the need of medical field like more focus on the privacy aspect.

For our case, we re-grouped specific metrics. In three main categories, we added some metrics that might be generally used and more relevant in the industrial domain and presented universal metrics. We propose the following three high-level abstract categories: representativeness, privacy, and realism, which will be elaborated and detailed in the following sub-section. We focused on metrics that can easily be applicable in real-world scenarios, utilizing the available tools for synthetic tabular data like SDV library and SynthEval, which data scientists in professional and industrial fields can use.

4.1 Representativeness

Representativeness is a critical measure that assesses how well the synthetic data is similar to the real data and how well the SDG can capture the underlying distributions, relations, and patterns in the real data. Many metrics can be used to assess this aspect.

4.1.1 Similarity

Statistical similarity is the simplest way to evaluate the quality of the synthetic data by comparing statistics like mean, median, and standard deviation. Missing value similarity is another metric that calculates the missing proportion of missing values in both real and synthetic data. Column shape similarity normalized the category column frequencies, then applied the chi-square χ^2 test to determine if the synthetic data comes from the same distribution as the real data. Another set of metrics tried to find the adherence and coverage of synthetic data to the real data rules; some metrics are Boundary adherence and Range coverage checks if the min. max. Values and range are bound to the real data values. Category adherence and Category coverage check if unique categories have the same proportion or presence in real and synthetic data. The frequency of data features compares the most frequent input features in both datasets to ensure that the frequency distribution is preserved, maintaining essential patterns in the data.

4.1.2 Distance based approaches

The distance-based approach is another set of metrics that focuses on measuring the proximity of probability distribution of real and synthetic data. Kolmogorov-Smirnov (KS) complement [66] uses the KS score applied to the cumulative distribution function (CDF) commonly used for continuous variables. Total variation distance (TVD) [67] applied to categorical values, checks the categorical probability distributions between the datasets; it first starts from the frequency of each category and then finds the probability then compares the probabilities between both pairs. Kullback-Leibler (KL) divergence (relative entropy) [68], is a prevalent metric to capture the distance between two data distributions and expressed by the following formula:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{1}$$

Where `P(x)` and `Q(x)` represent the real data and synthetic data probability, respectively, this metric is asymmetrical, meaning it measures how well `P(x)` approximates `Q(x)` and not the inverse; in our context, it signifies the amount of information loss that occurred to generate the synthetic data. Jensen-Shannon (JS) divergence [69] is a symmetrical version of KL divergence and produces a finite value, which is essential when making comparisons between multiple models. Hellinger distance [70] quantifies the Euclidean distance between the two distributions; this metric is easy to interpret as it ranges from 0, meaning identical distribution, to 1, signifying dissimilar distributions. Finally, Wasserstein distance [71] tries to estimate an observation as

synthetic or not; it incorporates both cumulative distributions and distance between points, making it more resilient to shape mismatches or outliers.

4.1.3 Correlation based approaches

Pairwise correlation distance (PCD) [72] measures the difference in a linear correlation between two pairs of variables from the datasets to check how well the relationship is preserved variable-wise. Smaller values indicate that the relationship is preserved. Correlation similarity indicates the same but on the level of the whole dataset instead of pairwise correlation and calculated on the correlation matrices of the data. Contingency similarity is applied to the categorical values by comparing the contingency table (similar to the confusion matrix but used to evaluate association rules); the interpretation of this metric is similar to the ones in this category.

4.1.4 Likelihood fitness

Data likelihood describes how likely synthetic data is to follow the same distribution as the real data. Metrics under this category fit a probabilistic model like a Bayesian network or Gaussian Mixture Models (GMMs) on the real data, then it evaluates how well the synthetic data will fit in the learned distribution. Higher likelihood implies that the synthetic data follows the same distribution and structure as the original data. GMMs likelihood learn the distribution as a mixture of multiple Gaussian distributions and applied for each row, it produces a value between $+\infty$, indicating that the row is likely part of the data, and $-\infty$, meaning it is not; this metric is commonly used for numerical data. Because of its interpretation complexity and to evaluate categorical data Bayesian network likelihood [73], it can be used, as it represents the dependencies between variables as a directed acyclic graph (DAG) representing relationships between variables. The range of this metric is between 0 and 1, 1 meaning the synthetic data is part of the real data.

4.1.5 Nearest neighbor adversarial

Nearest neighbor adversarial accuracy and resemblance loss measure how well an adversarial classifier model can differentiate between real and synthetic data based on nearest neighbors [74]. Implementing this metric, like the one in SynthEval [75], uses a neural network to calculate the nearest neighbors, and the lower the value, the more similar the synthetic data is to the real. However, it also indicates overfitting problems of the data if the value is very low, which renders this metric also helpful in detecting overfitting in models.

4.2 Privacy

Preserving privacy is one of the most important aspects of modern applications. It is one of the main motivations behind using synthetic data in general and tabular data precisely. Privacy in this field refers to the measures taken to ensure that the generated data does not expose sensitive information and details about individuals, personal identifiers, or sensitive data. It is crucial to have proven methods to measure and

quantify the degree of privacy preservation. Many approaches and metrics exist in the area, ranging from metrics suggested and used in tabular data for the medical field [76], where preserving the security of personal identifiers from leaking into the generated data is of the utmost importance. The industrial domain is also affected by this area, especially in the transformation carried by "Industry 4.0" standards, where the need for more simulation environments and machine learning models is rising as well with the need for more data to train those models, preserving and anonymizing these data before being used in downstream tasks is required [77]. Metrics for measuring the level of privacy preservation are essential to provide a quantitative way to assess whether the synthetic data obscures sensitive information while at the same time keeping the utility and usability of the overall data.

4.2.1 Correct attribution probability

Correct attribute probability (CAP) is a privacy metric to evaluate the disclosure risk of sensitive information from the synthetic data introduced and elaborated in [78–80]. It assumes an attacker/adversary combines the knowledge of some non-sensitive attributes in the real data with the synthetic data to attribute the sensitive parts of the real data correctly. CAP calculates the probability of finding the correct sensitive target based on the provided vital attributes. In practice, CAP is calculated as the probability of a target attribute's value appearing in the records with the same essential attributes. A higher occurrence of the target value in the records means a higher probability of the adversary predicting the target attribute correctly. This metric can be used directly for categorical values from the SDMetrics library by specifying the sensitive and non-sensitive parts of the real dataset and then combining them with the obtained synthetic data.

4.2.2 Privacy against inference

Privacy against inference is another sub-group of metrics that can assess the probability of correctly attributing sensitive data from a constructed knowledge of non-sensitive fields in real data and complete synthetic data. The preliminary construction of the initial dataset is the same as that of CAP. However, instead of finding the probability of target attributes appearing in the records, it extends a machine learning algorithm to classify a categorical or predict a numerical value. Multi-layer perception, KNN, ensemble learning, and other machine learning algorithms can be trained on the constructed dataset.

4.2.3 Novelty

Novelty is an indirect way to measure privacy; in this methodology, we try to check if a given row from synthetic data is entirely new and not duplicating data in real data. Many measures under this category can be used. New row synthesis (NRS) is a simple metric where we check each synthetic row against all the real rows to check if the attributes inside are identical to the original dataset. An exact match can be used for categorical rows, while for numerical data, we can estimate the proximity to normalized value in the data range. Distance to closest record (DCR) is another metric

used in [81, 82], to assess how close a synthetic record is to real records. The distance measure can be Euclidean distance for a numerical data point or Hamming distance for categorical data. When the distance to the nearest record is minimal, it indicates that the model failed to generate synthetic data. At the same time, it indicates privacy concerns of possible data leakage. Max real to synthetic similarity (Max-RTS) is a measure that can be used to understand if the model can generate new records and not only duplicate real data [83]. In the privacy context, it can also signify how private our new data is. The working of this metric is similar to that of DCR as they both rely on the distance between records. Maximum mean discrepancy MMD proposed by [84] is a statistical-based kernel metric that finds the similarity between two distributions; it involves embedding data points from real data and synthetic data into high-dimension feature space using kernel function and then calculating the difference between the two distribution. A higher MMD value indicates more novelty and privacy but risks the model not adhering to real data distribution.

4.3 Realism

Data realism can also be referred to as application fidelity, which is the set of metrics that decides whether the generated data is a replacement for the real data in downstream tasks like training ML models. The realism metrics ensure that the generated data resembles the original data in terms of distribution and data and can achieve similar performance when, for example, used to train ML models. The general validation form starts with a set of synthetic data and uses this set to train an ML model. Afterward, another ML model can be trained on the real data, and finally, evaluation can be done on a separate evaluation subset of the real data. Performance can be evaluated using accuracy, precision, F1-score, R^2, and other machine learning evaluation metrics. Usually, ML models can include MLP, SVM, regression models, ensemble learning, and decision trees. This approach is widely used as a benchmark for any model used for SDG tasks. Additionally, other metrics like propensity mean-squared-error (pMSE) can be used [85]; the core idea behind pMSE is to measure if a machine learning model can distinguish between real and synthetic data, in optimal cases if the real and synthetic data are indistinguishable it means that the model succeeded in learning the underlying relation between the data, and can usable data.

4.4 Global metrics

Global metrics are a set of metrics that evaluate multiple criteria simultaneously, like checking if the data is realistic, private, and representative. TabSynDex is a proposed global metric in [86] as a single interpretable score that can assess the quality of the synthetic data based on multiple dimensions. The dimensions covered are: (i) statistical measures including mean, median, and standard deviation, (ii) correlation score of the categorical and continuous columns in both datasets, (iii) pMSE score for realism, (iv) support coverage to check whether minority categories found in the real data are represented in the synthetic version, and (v) machine learning efficiency were multiple machine learning models are used for the evaluation, root mean square error (RMSE) is used for regression cases, while F1-score is used for classification cases. Each

dimension score is calculated to penalize bad examples; in the case of the correlation score, log transformation is used to reduce the impact of small correlation values. Ultimately, all the individual scores are aggregated into one score, ranging between 0, meaning the data differs from the real data in all dimensions, and 1, meaning the data is complementary in all the dimensions. It is worth mentioning that TabSynDex attributes are of equal importance for each dimension.

Three-dimensional metric proposed in [87], consists of (i) $\alpha - precision$ representing the realism dimension, which is calculated by finding the smallest subset of the real data that contains (α) portion of the data, and then checking the synthetic data distribution falls under the same portion, (ii) $\beta - recall$ evaluates how well the synthetic data covers the range of samples in the real data, (iii) authenticity or generalization which checks if the samples are new or copy from the real data. This metric is important because it provides a more specific indication of the dataset. In contrast, TabSynDex provides one number; it can be challenging to interpret the problematic dimension. A three-dimensional metric can provide a better indication of each dimension depending on the acceptable trade-off in the application.

5 Experiments

5.1 Experiment setup

All the experiments were realized on an Nvidia DGX server equipped with 8xNvidia A100 GPUs, each with 40 GB of memory. The high-performance infrastructure allowed multiple training sessions simultaneously, and evaluation runs used the same hardware. The datasets used were publicly available dataset sources from the UCI machine learning repository, Kaggle, unlearn for Bayesian simulated datasets, and Yann-lecun for the Mnist28 dataset. For the models' implementation, we either sourced them from the original GitHub implementation proposed in their papers, or for synthetic data vault (SDV)-related models, we used SDV directly for the dataset processing, training, and sampling as it offers numerous tools and ready-to-use functions for this reason. The evaluation was done using the implementation of the mentioned metrics from different libraries like SDMetrics, SynthEval, Pomegranate, and sci-kit-learn, which were explicitly used for machine learning efficiency. Concerning the experiment protocol, we conducted a total of three runs for each experiment, and the evaluation was repeated three times; the results for each metric were later averaged out to minimize variability.

5.2 Metrics selection

Evaluation metrics used in our experiments are picked from Section 4. We narrowed the selection to the more significant ones that can give better conclusions when inspected. We present the metrics used as shown in Table 2. The metrics are from the three proposed categories: representativeness, privacy, and realism. For representativeness, we chose nearest neighbor adversarial accuracy and likelihood fitness because they provide a more complex understanding of the selected category than the more straightforward metrics that measure common statistical properties of the dataset. It is worth

Table 2: Summary of metrics used for the evaluation, with each respective category and the source implementation used.

Metric	Category	Source
Nearest Neighbour Adversarial Accuracy(↓)$^{\odot}$	Representativeness	SynthEval
Bayesian Likelihood Fitness(↑)*	Representativeness	SDMetrics
Gaussian Likelihood Fitness(↑)*	Representativeness	SDMetrics
Propensity Mean Squared Error(↓)†	Realism	SynthEval
Decision Tree(↑)$^{\odot\diamond}$	Realism	scikit-learn
Random Forest Classifier(↑)$^{\odot\diamond}$	Realism	scikit-learn
Multi-layer Perceptron (MLP) Classifier(↑)$^{\odot\diamond}$	Realism	scikit-learn
Logistic Regression(↑)$^{\odot\diamond}$	Realism	scikit-learn
Multinomial Logistic Regression(↑)$^{\odot\diamond}$	Realism	scikit-learn
Support Vector Machine (SVM)(↑)$^{\odot\diamond}$	Realism	scikit-learn
New Row Synthesis(↑)$^{\odot}$	Privacy	SDMetrics
Privacy Against Inference(↑)$^{\odot}$	Privacy	SDMetrics
Correct Attribution Probability(↑)$^{\odot}$	Privacy	SDMetrics

(↑): higher score is better; (↓): lower score is better; ($\odot$): score ranges between [0,1]; ($\star$): value ranges between [$-\infty$ or 0, $+\infty$ or 1]; ($\dagger$): value ranges between [0,0.25]; ($\diamond$): used for classification target dataset, F1-Score is reported

mentioning that for likelihood fitness, Gaussian Mixture is used for continuous values, while Bayesian likelihood is used for categorical data. We chose various models most commonly used in ML tasks in the realism category, especially in the industrial domains, because of their proven performance. We chose metrics representing multiple subcategories for privacy, covering novelty, privacy against inference, and correct attribution probability.

5.3 Datasets selection

We opt to use public and open-source datasets found in the general domain. The primary motivation behind this decision is to allow the reproducibility of the results. Additionally, most research in the domain of tabular data generation uses one of those datasets as a benchmark for their results. We also argue that the selected dataset covers many versions and types of datasets found in other domains, including the industrial domain, where the dataset target can be classification or regression. Moreover, the selected dataset encapsulates different variable types, from continuous values, categorical values, and mixed-type to textual data. For the classification problem, we used Census, Covtype, Child, Adult, and Health, while for the regression problems, we used Fake Hotel Guests, Insurance, Expedia Hotel Logs, and News.

5.4 Models selections

The model selection was based on each category from the proposed taxonomy, and they represent the state-of-the-art models in terms of adoption and impact in recent studies. For the statistical models, we chose Gaussian coupla based on the SDV implementation. This model is also able to capture complex dependencies in tabular data. We used CTGAN, TVAE, and CopulaGAN from the SDV in the GAN-based category. These models can generate synthetic data and capture different types of data distributions. They are used as a comparison baseline for other models' respective evaluations.

We include TabDDPM, a model based on the diffusion process, a process widely used in other generative tasks like image generation. For the implementation of TabDDPM, we followed the published implementation by the authors. Finally, GReat and Tabula are emerging models based on LLMs, and they represent the advancements in integrating transformer architecture for tabular data generation. All the selected models represent different advancements in each category and can be considered the baseline for evaluating any new model targeting the tabular generation field.

5.5 Experimentation and results

In this section, we will present the experimentation done, with a general overview of the parameters used and the evaluation process. For the census and CovType datasets, GAN and Gaussian Coupla were trained on 150 epochs, LLM-based models were trained for 50 epochs, and TabDDPM for 5000 epochs to guarantee the best convergence for the model. The resulting models were sampled to 50K records for the evaluation process. However, we faced issues with TabDDPM not converging and MLoss staying at 0 all the time. In the sampling process, the model could not handle NaN cases, the same scenario happened when using other complex datasets like Fake Hotel Guests, Health, and Expedia Hotel Logs. Similar issues were faced with Tabula during the sampling process for Census, Fake Hotel Guests, Adults, Health, and News, where the trained model could not sample any points. We tried repeating the experiments multiple times, with different parameter selections for the training and sampling, but it did not work. For both issues, we tried reaching the authors regarding the published implementation but did not get a response until publishing; both models will be excluded from the mentioned datasets. The fake Hotel Guests dataset was trained for 400 epochs on GAN models and Gaussian Copula and 200 epochs for the GReat model using distilgpt-2 for all the experiments. In the Insurance, Child, Adult, and News experiments, the training for GAN and Gaussian Copula was for 150 epochs, and LLM models were trained for 50 epochs and TabDDPM for 4096 epochs. The health dataset experiment used 150 epochs for the GAN and Gaussian Copula models and 75 epochs for the GReat model; the Expedia Hotel Logs experiment needed more training for convergence, so we ran all the models for 300 epochs. We retried each experiment three times, with three resulting sampled datasets, and reported the metrics with the same sample real dataset; for metrics requiring the selection of sensitive fields, we chose personal or risky columns and gave the rest to the evaluation algorithm to build a combined dataset. The results from representativeness results are shown in Table 4, privacy results are shown in Table 3, and realism are shown in Table 6.

6 Results Discussion

In this section, we discuss the evaluation results. The privacy metrics across various models and public datasets reveal significant insights, as shown in Table 3. Most models achieve high scores in NRS and PAI, indicating the generation of unique data. However, the CAP metric raises concerns across all datasets, except for Insurance and Child. For most datasets, CAP values close to 0 suggest the potential risk that

Table 3: Privacy evaluation of public dataset over selected models. Used metrics: New Row Synthesis (NRS), Privacy Against Interference (PAI), Correct Attribution Probability (CAP).

	Census			Covtype			Fake Hotel Guests			Insurance			Child			Adult			Health			Expedia Hotel Logs			News		
	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP	NRS	PAI	CAP
Gaussian Copula	1	1	0	1	1	0	1	1	0	0.999	1	1	0.994	0.998	0.999	1	0.999	0	1	1	0	1	1	0	1	1	0
CTGAN	1	0.999	0	1	1	0	1	1	0	1	1	1	0.999	1	0.999	1	0.987	0	1	1	0	1	1	0	1	1	0
TVAE	1	0.986	0	1	1	0	1	1	0	0.948	0.999	0.999	0.862	0.991	0.996	0.999	1	0	0.999	1	0	1	1	0	1	1	0
CopulaGAN	1	1	1	1	1	0	1	1	0	0.983	1	0.999	0.918	0.987	0.996	0.999	0.965	0	1	1	0	1	1	0	1	1	0
Tabula	-	-	-	1	1	0	-	-	-	-	1	0	1	1	1	-	-	-	-	-	-	1	1	0	-	-	-
BeGReaT	0.626	0.618	0.328	1	1	0	1	1	0	0.792	1	0.999	0.577	0.985	0.989	0.998	0.938	0	1	1	1	1	1	0	1	1	0
TabDDPM	-	-	-	1	0.999	0	-	-	-	0.958	1	0.999	0.932	0.993	0.998	0.999	0.976	0	-	-	-	-	-	-	1	1	0

Table 4: Representativeness evaluation of public dataset over selected models used the metrics: Gaussian Likelihood Fitness (GM), and Nearest Neighbor Adversarial Accuracy (NNAA).

	Census		Covtype		Fake Hotel Guests		Adult		Health		Expedia Hotel Logs		News	
	GM	NNAA	GM	NNAA	GM	NNAA	GM	NNAA	GM	NNAA	GM	NNAA	GM	NNAA
Gaussian Copula	-1.10×10^9	0.934	-2.56×10^5	0.999	-4.17×10^6	0.892	-1.04×10^{12}	0.837	-181	0.84	-3.90×10^4	0.877	-195	0.983
CTGAN	-7.77×10^8	0.938	-2.97×10^5	0.949	$\mathbf{-4.09 \times 10^6}$	**0.55**	-2.87×10^9	0.815	-234	0.921	-8.29×10^4	0.827	-745	0.906
TVAE	-7.32×10^4	0.885	-3.10×10^4	0.845	-9.03×10^6	0.605	-3.56×10^9	0.747	-147	**0.85**	**-993**	0.904	-131	0.952
CopulaGAN	-2.16×10^9	0.852	-3.37×10^5	0.999	-6.30×10^6	0.865	-9.06×10^9	0.663	-202	0.899	-2.13×10^4	0.87	-171	0.989
Tabula	-	-	-4.28×10^{10}	0.798	-	-	-	-	-	-	-2602	0.838	-	-
BeGReaT	**21**	**0.575**	**-1238**	0.868	-4.61×10^8	0.717	**-9**	0.594	**-92**	0.879	-3570	**0.73**	**-124**	0.703
TabDDPM	-	-	-2250	**0.589**	-	-	-4.40×10^7	**0.568**	-	-	-	-	-9.60×10^5	**0.635**

Table 5: Representativeness evaluation of public dataset over selected models used the metrics: Bayesian Likelihood Fitness (BLF), and Nearest Neighbor Adversarial Accuracy (NNAA).

	Insurance		Child	
	BLF	NNAA	BLF	NNAA
Gaussian Copula	2.14×10^{-7}	0.835	2.59×10^{-6}	0.618
CTGAN	4.27×10^{-10}	0.838	3.23×10^{-7}	0.728
TVAE	6.14×10^{-7}	0.788	4.79×10^{-5}	0.449
CopulaGAN	2.55×10^{-7}	0.834	6.08×10^{-5}	0.502
Tabula	0	1	0	0.999
BeGReaT	$\mathbf{1.85 \times 10^{-5}}$	**0.676**	$\mathbf{4.17 \times 10^{-4}}$	**0.372**
TabDDPM	3×10^{-7}	0.750	2.65×10^{-5}	0.392

Table 6: Realism evaluation of public dataset over selected models used. Used metrics: F1-Score and Propensity Mean Squared Error (PMSE).

	Census		Covtype		Child		Adult		Health	
	F1-score	PMSE	F1-score	PMSE	F1-score	PMSE	F1-score	PMSE	F1-score	PMSE
Gaussian Copula	0.515	0.186	0.116	0.190	0.500	0.024	0.174	0.096	0.727	0.052
CTGAN	0.508	**0.033**	0.091	0.031	0.485	0.042	0.151	0.010	0.484	0.052
TVAE	0.655	0.040	0.311	0.034	0.626	0.039	0.265	0.028	0.908	**0.043**
CopulaGAN	0.651	0.045	0.102	0.153	0.596	0.037	0.266	0.005	0.851	0.044
Tabula	-	-	0.198	0.092	0.289	0.119	-	-	-	-
BeGReaT	**0.691**	0.092	0.289	0.099	**0.636**	0.052	**0.286**	0.021	**0.958**	0.171
TabDDPM	-	-	**0.370**	**0.001**	0.616	**0.016**	0.281	**0.0001**	-	-
Real	0.662	-	0.450	-	0.629	-	0.302	-	0.970	-

adversaries could access non-sensitive fields in the real dataset and, by combining this with the synthetic data, accurately infer sensitive values from the original dataset.

The representativeness evaluation is shown in Table 4 and 5, for the categorical datasets in the first table, BeGReat demonstrated the highest performance for GM metric across datasets like Census, Covtype, Adult, Health, and News, which indicates a solid ability to generate realistic data, while also achieving the lowers NNAA score, particularly in Census and Expedia Hotel Logs. CTGAN also performed well on Fake Hotel Guests with the best GM and NNAA. TVAE exhibited good performance for NNAA in the Health dataset and the highest GM for Expedia Hotel Logs. Aside from the mentioned models, Gaussian Copula and CopulaGAN showed poor performance across most datasets compared to the other models, indicating less representative data. For the numerical datasets, BeGReat outperformed all other insurance and child datasets models in both metrics, BLF and NNAA. Overall, BeGReat and CTGAN were the top performers when generating data similar to the original data.

For the realism metric as shown in Table 6, the evaluation was done based on F1-Score for datasets with classification target and PMSE. The F1-Score was reported based on the average of the F1-Score using a decision tree, random forest, multi-layer perceptron classifier, logistic regression, multinomial regression, and support vector

machine SVM. The evaluation results showed that BeGReat performed best in F1-Score on Census, Child, Adult, and Health datasets. Another significant interpretation can be observed in the F1-Score for models trained on Census and Child, which achieved better results than the baseline real dataset. This can be because LLMs can understand the relation between variables more deeply and generate records that might be more significant for the models' training. TabDDPM also performed well, especially for PMSE for all datasets where generating synthetic data was possible. This indicates that the model resembles real data statistically and in terms of propensity scores. On the other hand, Gaussian Copula showed moderate F1-Score in some datasets but generally had higher PMSE values, which indicates less realistic data than the others. CTGAN performed well for some datasets like Census and achieved the best PMSE for the same dataset. Overall, BeGreat and TabDDPM performed better than the other models, demonstrating their effectiveness in producing realistic synthetic data. At the same time, CTGAN, TVAE, and CopulaGAN showed more moderate capabilities but still excelled in some datasets.

In conclusion, the evaluation results highlight the BeGReat and TabDDPM models as the best performers across different dimensions and different datasets. Nevertheless, we showed the importance of evaluating each dimension and assessing different metrics that tell different parts of the story. Additionally, GAN-based models showed good results for different use cases, while more statistical models showed lower performance. In other words, LLM and diffusion-based approaches are now emerging in the field of synthetic data generation, while GAN models can still be competitive in multiple situations. Moreover, we faced multiple problems with Tabula while training and generating data based on a more numerical dataset, which should be considered when using such a model in cases where the dataset contains numerical columns. We also faced issues with TabDDPM as it was unable to handle missing values in the training dataset and while generating records.

7 Conclusion

In this paper, we tried to survey the state-of-the-art techniques and models for synthetic tabular data generation focused on industrial applications. However, the findings in this paper can be applied to other domains, especially the evaluation dimensions we proposed. To the best of our knowledge, existing surveys in the field either focus only on tabular data generation in the medical domain or only study the usage of GAN-based models for SDG. On the other hand, existing studies for evaluation metrics try to explore multiple evaluation metrics and group them under different categories. We built upon existing work and studied individual metrics while reorganizing them into multiple dimensions.

In our study, we have shown recent generation models that are not only limited to GAN but also extend to the integration of diffusion models that previously were used for image generation tasks into tabular generation tasks. LLMs are also being integrated into this field, and future trends look promising because of the internal capabilities of LLMs to capture semantic information based on learned knowledge through their foundation models. Public datasets are also important for benchmarking any

existing and new model; we presented multiple public datasets from different domains that can be used to validate models. In the evaluation section, we discussed and proposed a taxonomy that evaluates a model on multiple dimensions with higher-level categories like realism, which checks the utility of the dataset in certain downstream applications, representativeness to check if the synthetic data is compliment with the original dataset, and privacy that measures any data leakage for sensitive information, which is at utmost importance for different domains, especially industrial domain. Under each category, we proposed multiple subcategories and metrics to measure different aspects of the data, so we can simulate and understand the behavior of the synthetic data when used as a replacement or augmentation to real data. Finally, we conducted multiple evaluations and experiments to test models under each section of the taxonomy and showcase their usability under different datasets. The evaluation concluded by showing how important it is to measure a model's generation capabilities under different metrics, for example, in privacy settings, using NRS and PAI showed most models as safe to use, while using CAP metrics raised concerns in certain models and datasets. Over and above that, we showed the exceptional performance of newly emerging models like BeGReat based on LLMs and TabDDPM based on diffusion models in tabular data generation, and this can fuel further research into adopting LLMs and diffusion models for tabular data generation while also resolving some common issues faced with such models like handling numerical data, missing data, and how to effectively encode data like numerical ones into a format suitable for transformer to work with.

In light of the findings from this study, several points can be addressed in future work. The first direction is exploring the integration of differential privacy in generative models to address some challenges with low CAP scores observed in some datasets and check if case-specific privacy mechanisms can be applied to overcome privacy challenges. As mentioned, to obtain the best data, one model can only fit some use cases, and here, the importance of more specialized models and techniques for data processing should be addressed in future studies. Another aspect concerning new models like LLMs should be improved to overcome challenges with understanding and handling numerical data, managing flaws in the datasets like missing values and outliers, and reducing training time for more practical solutions. With the revolution in foundation models like GPT and LLAMA, more work should be done to study and adapt such models for tabular data generation, especially after the proven capabilities encapsulated in those models. Finally, developing and incorporating new evaluation metrics that capture more fine aspects of privacy, realism, and representativeness can ensure highly private and valuable data for further tasks. These future work and directions can solve multiple impediments in this field, making tabular synthetic datasets more used and adopted in all domains, specifically the industrial domain.

Declarations

Funding. This research is supported by the EIPHI Graduate School (contract "ANR-17-EURE-0002") and the BMW TechOffice Munich.

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

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. Our generated synthetic datasets and experimentation are available from the corresponding authors upon reasonable request.

Materials availability. Not applicable.

Code availability. The code used for training and evaluation is available from the corresponding authors upon reasonable request.

Author contribution. Not applicable.

References

[1] Schuh, G., Anderl, R., Gausemeier, J., Ten Hompel, M. & Wahlster, W. *Industry 4.0 maturity index: shaping the digital transformation of companies* (Herbert Utz Verlag, 2017).

[2] Zuehlke, D. Smartfactory—towards a factory-of-things. *Annual reviews in control* **34**, 129–138 (2010).

[3] Commission, E. *et al. Industry 5.0 – Towards a sustainable, human-centric and resilient European industry* (Publications Office of the European Union, 2021).

[4] Xu, X., Lu, Y., Vogel-Heuser, B. & Wang, L. Industry 4.0 and industry 5.0—inception, conception and perception. *Journal of manufacturing systems* **61**, 530–535 (2021).

[5] Sajid, S. *et al.* Data science applications for predictive maintenance and materials science in context to industry 4.0. *Materials today: proceedings* **45**, 4898–4905 (2021).

[6] Makkar, S., Devi, G. N. R. & Solanki, V. K. *Applications of machine learning techniques in supply chain optimization*, 861–869 (Springer, 2020).

[7] Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* **81**, 84–90 (2022).

[8] Sukhobok, D., Nikolov, N. & Roman, D. *Tabular data anomaly patterns*, 25–34 (IEEE, 2017).

[9] Espinosa, E. & Figueira, A. On the quality of synthetic generated tabular data. *Mathematics* **11**, 3278 (2023).

[10] Das, H. P. *et al. Conditional synthetic data generation for robust machine learning applications with limited pandemic data*, Vol. 36, 11792–11800 (2022).

[11] Abou Akar, C. *et al.* Generative adversarial network applications in industry 4.0: A review. *International Journal of Computer Vision* **132**, 2195–2254 (2024).

[12] Figueira, A. & Vaz, B. Survey on synthetic data generation, evaluation methods and gans. *Mathematics* **10**, 2733 (2022).

[13] Fonseca, J. & Bacao, F. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data* **10**, 115 (2023).

[14] Borisov, V. *et al.* Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems* (2022).

[15] Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A. & Zahariadis, T. A review of tabular data synthesis using gans on an ids dataset. *Information* **12**, 375 (2021).

[16] 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).

[17] Pezoulas, V. C. *et al.* Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal* (2024).

[18] Sahakyan, M., Aung, Z. & Rahwan, T. Explainable artificial intelligence for tabular data: A survey. *IEEE access* **9**, 135392–135422 (2021).

[19] Došilović, F. K., Brčić, M. & Hlupić, N. *Explainable artificial intelligence: A survey*, 0210–0215 (2018).

[20] Fang, X. *et al.* Large language models (llms) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research* (2024). URL https://www.amazon.science/publications/large-language-models-llms-on-tabular-data-prediction-generation-and-understanding-a-survey.

[21] Yin, P., Neubig, G., Yih, W.-t. & Riedel, S. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314* (2020).

[22] Herzig, J., Nowak, P. K., Müller, T., Piccinno, F. & Eisenschlos, J. M. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349* (2020).

[23] Heckerman, D. Bayesian networks for data mining. *Data mining and knowledge discovery* **1**, 79–119 (1997).

[24] Kamthe, S., Assefa, S. & Deisenroth, M. Copula flows for synthetic data generation. *arXiv preprint arXiv:2101.00598* (2021).

[25] Goodfellow, I. *et al.* Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).

[26] Arjovsky, M., Chintala, S. & Bottou, L. *Wasserstein generative adversarial networks*, 214–223 (PMLR, 2017).

[27] Mirza, M. & Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[28] Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*

arXiv:1511.06434 (2015).

[29] Xu, L. & Veeramachaneni, K. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264* (2018).

[30] Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling tabular data using conditional gan. *Advances in neural information processing systems* **32** (2019).

[31] Uci machine learning repository. URL https://archive.ics.uci.edu/. Accessed: 2024-07-14.

[32] Park, N. *et al.* Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384* (2018).

[33] Zhao, Z., Kunar, A., Birke, R. & Chen, L. Y. *Ctab-gan: Effective table data synthesizing*, 97–112 (PMLR, 2021).

[34] Zhao, Z., Kunar, A., Birke, R., Van der Scheer, H. & Chen, L. Y. Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data* **6**, 1296508 (2024).

[35] Wen, B., Cao, Y., Yang, F., Subbalakshmi, K. & Chandramouli, R. *Causal-tgan: Modeling tabular data using causally-aware gan* (2022).

[36] Ho, J., Jain, A. & Abbeel, P. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds) *Denoising diffusion probabilistic models.* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) *Advances in Neural Information Processing Systems*, Vol. 33, 6840–6851 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[37] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. *Deep unsupervised learning using nonequilibrium thermodynamics*, 2256–2265 (PMLR, 2015).

[38] Kotelnikov, A., Baranchuk, D., Rubachev, I. & Babenko, A. *Tabddpm: Modelling tabular data with diffusion models*, 17564–17579 (PMLR, 2023).

[39] Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds) *Revisiting deep learning models for tabular data.* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) *Advances in Neural Information Processing Systems*, Vol. 34, 18932–18943 (Curran Associates, Inc., 2021). URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf.

[40] Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[41] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[42] Minaee, S. *et al.* Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[43] Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[44] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[45] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).

[46] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

[47] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**, 1–67 (2020).

[48] Lewis, M. *et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[49] Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M. & Kasneci, G. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280* (2022).

[50] Becker, B. & Kohavi, R. Adult. UCI Machine Learning Repository (1996). DOI: https://doi.org/10.24432/C5XW20.

[51] Wallace, E., Wang, Y., Li, S., Singh, S. & Gardner, M. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940* (2019).

[52] Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

[53] Kaggle. Kaggle: Your machine learning and data science community (2024). URL https://www.kaggle.com/. Accessed: 2024-07-14.

[54] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[55] Bellman, R., Bellman, R. & Corporation, R. *Dynamic Programming* Rand Corporation research study (Princeton University Press, 1957). URL https:

//books.google.de/books?id=rZW4ugAACAAJ.

[56] Zhao, Z., Birke, R. & Chen, L. Tabula: Harnessing language models for tabular data synthesis. *arXiv preprint arXiv:2310.12746* (2023).

[57] Nash, S. T. T. S. C. A., Warwick & Ford, W. Abalone. UCI Machine Learning Repository (1994). DOI: https://doi.org/10.24432/C55C7W.

[58] Blackard, J. Covertype. UCI Machine Learning Repository (1998). DOI: https://doi.org/10.24432/C50K5N.

[59] Binder, J., Koller, D., Russell, S. & Kanazawa, K. Adaptive probabilistic networks with hidden variables. *Machine Learning* **29**, 213–244 (1997).

[60] Adam, W. K. Expedia hotel recommendations (2016). URL https://kaggle.com/competitions/expedia-hotel-recommendations.

[61] Singh, K. Facebook Comment Volume. UCI Machine Learning Repository (2015). DOI: https://doi.org/10.24432/C5Q886.

[62] Stolfo, F. W. L. W. P. A., Salvatore & Chan, P. KDD Cup 1999 Data. UCI Machine Learning Repository (1999). DOI: https://doi.org/10.24432/C51C7N.

[63] Dankar, F. K., Ibrahim, M. K. & Ismail, L. A multi-dimensional evaluation of synthetic data generators. *IEEE Access* **10**, 11147–11158 (2022).

[64] Dankar, F. K., Ibrahim, M. K. & Ismail, L. A multi-dimensional evaluation of synthetic data generators. *IEEE Access* **10**, 11147–11158 (2022).

[65] Goncalves, A. *et al.* Generation and evaluation of synthetic patient data. *BMC medical research methodology* **20**, 1–40 (2020).

[66] Arora, S., Ge, R., Liang, Y., Ma, T. & Zhang, Y. *Generalization and equilibrium in generative adversarial nets (gans)*, 224–232 (PMLR, 2017).

[67] Tao, L., Xu, S., Wang, C.-H., Suh, N. & Cheng, G. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv preprint arXiv:2405.15337* (2024).

[68] Dandekar, A., Zen, R. A. & Bressan, S. *Comparative evaluation of synthetic data generation methods* (2017).

[69] Arora, S., Ge, R., Liang, Y., Ma, T. & Zhang, Y. *Generalization and equilibrium in generative adversarial nets (gans)*, 224–232 (PMLR, 2017).

[70] Le Cam, L. M. & Yang, G. L. *Asymptotics in statistics: some basic concepts* (Springer Science & Business Media, 2000).

[71] Robinson, A. & Turner, K. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology* **1**, 241–261 (2017).

[72] Khrulkov, V. & Oseledets, I. *Geometry score: A method for comparing generative adversarial networks*, 2621–2629 (PMLR, 2018).

[73] Ping, H., Stoyanovich, J. & Howe, B. *Datasynthesizer: Privacy-preserving synthetic datasets*, 1–5 (2017).

[74] Yale, A. *et al. Synthesizing quality open data assets from private health research studies*, 324–335 (Springer, 2020).

[75] Lautrup, A. D., Hyrup, T., Zimek, A. & Schneider-Kamp, P. Syntheval: A framework for detailed utility and privacy evaluation of tabular synthetic data. *ArXiv* **abs/2404.15821** (2024). URL https://api.semanticscholar.org/CorpusID:269362787.

[76] Yale, A. *et al. Privacy preserving synthetic health data* (2019).

[77] Ogrezeanu, I. *et al.* Privacy-preserving and explainable ai in industrial applications. *Applied Sciences* **12**, 6395 (2022).

[78] Elliot, M. Final report on the disclosure risk associated with the synthetic data produced by the sylls team. *Report 2015* **2** (2015).

[79] Taub, J., Elliot, M., Pampaka, M. & Smith, D. *Differential correct attribution probability for synthetic data: an exploration*, 122–137 (Springer, 2018).

[80] Hittmeir, M., Mayer, R. & Ekelhart, A. *A baseline for attribute disclosure risk in synthetic data*, 133–143 (2020).

[81] Park, N. *et al.* Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11**, 1071–1083 (2018). URL https://doi.org/10.14778/3231751.3231757.

[82] Sivakumar, J., Ramamurthy, K., Radhakrishnan, M. & Won, D. Generativemtd: A deep synthetic data generation framework for small datasets. *Knowledge-Based Systems* **280**, 110956 (2023). URL https://www.sciencedirect.com/science/article/pii/S0950705123007062.

[83] Rashidian, S. *et al.* Michalowski, M. & Moskovitch, R. (eds) *Smooth-gan: Towards sharp and smooth synthetic ehr data generation.* (eds Michalowski, M. & Moskovitch, R.) *Artificial Intelligence in Medicine*, 37–48 (Springer International Publishing, Cham, 2020).

[84] Norgaard, S., Saeedi, R., Sasani, K. & Gebremedhin, A. H. *Synthetic sensor data generation for health applications: A supervised deep learning approach*, 1164–1167 (2018).

[85] Snoke, J., Raab, G. M., Nowok, B., Dibben, C. & Slavkovic, A. General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society Series A: Statistics in Society* **181**, 663–688 (2018). URL https://doi.org/10.1111/rssa.12358.

[86] Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M. & Narang, P. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence* **5**, 300–309 (2022).

[87] 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*, 290–306 (PMLR, 2022).